ARI Research Note 2009-02

Culturally Aware Agents for Training Environments (CAATE): Phase I Final Report

Scott Neal Reilly, Clare Bayley, David Koelle, Stephen Marotta, and Jonathan Pfautz,
Charles River Analytics

Michael Keeney,
Aptima, Inc.

Michael J. Singer
U.S. Army Research Institute

ARI-Orlando Research Unit
Stephen L. Goldberg, Chief

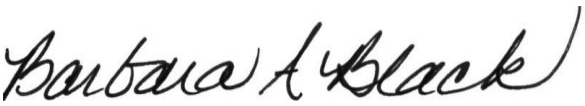
January 2009

U.S. Army Research Institute
for the Behavioral and Social Sciences

U.S. Army Research Institute for the Behavioral and Social Sciences

A Directorate of the Department of the Army Deputy Chief of Staff, G1

Authorized and approved for distribution:



BARBARA A. BLACK, Ph.D.
Research Program Manager
Training and Leader Development

MICHELLE SAMS, PhD.
Director

Research accomplished under contract
for the Office of the Secretary of Defense

Charles River Analytics, Inc. & Aptima, Inc.

Technical review by:

Dr. Christian Jerome, U.S. Army Research Institute

NOTICES

DISTRIBUTION: Primary distribution of this Research Note has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, Attn: DAPC-ARI-MS, 2511 Jefferson Davis highway, Arlington, Virginia 22202-3926.

FINAL DISPOSITION: This Research Note may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this Research Note are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE:	2. REPORT TYPE:	3. DATES COVERED:
January 2009	Final	April 2007 – November 2007

4. TITLE AND SUBTITLE:	5a. CONTRACTOR OR GRANT NUMBER:
Culturally Aware Agents for Training Environments (CAATE): Phase I Final Report	W91WAW-07-P-0250
	5b. PROGRAM ELEMENT NUMBER: 622785

6. AUTHOR(S):	5c. PROJECT NUMBER:
Scott Neal Reilly, Clare Bayley, David Koelle, Stephen Marotta, Jonathan Pfautz, Charles River Analytics, Inc., Michael Keeney, Aptima, & Michael J. Singer, ARI	A790
	5d. TASK NUMBER: 294
	5e. WORK UNIT NUMBER:

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER:
Charles River Analytics Inc. 625 Mount Auburn Street Cambridge, MA 02138	SBIR Topic #: OSD06-CR4 DoD Proposal #: O063-CR4-2094

9. SPONSORING / MONITORING ABENCY NAME(S) AND ADDRESS(ES):	10. MONITOR ACRONYM:
U.S. Army Research Institute for the Behavioral and Social Sciences 2511 Jefferson Davis Highway Arlington, VE 22202-3926	ARI
	11. MONITOR REPORT NUMBER: Research Note 2009-02

12. DISTRIBUTION / AVAILABILITY STATEMENT:
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES:
Contracting Officer's Representative and Subject Matter POC: Dr. Michael J. Singer

14. ABSTRACT *(Maximum 200 words)*:
Recently, the U.S. Army has conducted a wide range of missions within the context of very different cultures and languages. These missions often require junior leaders and Soldiers to interact and communicate effectively with people whose cultures, languages, lifestyles, and beliefs are very different from those found in the U.S. Computer-based training in virtual environments has the potential to train Soldiers to rehearse missions with a sound knowledge of the relevant local cultural context. Existing computer simulations of culturally situated agents representing humans are currently very limited in fidelity, making them unsuitable for training and rehearsal. This effort investigated, designed and demonstrated the feasibility of a two-step approach addressing the modeling of believable cultural agents. First, a mission essential competencies approach identifies key skills needed in training. Second, a modeling toolkit for designing computer-controlled agents for cultural training applications was described. The approach uses social network modeling technologies to develop models of interconnected agents within a graphical environment and a human behavior modeling tool for simulated agents based on the cultural context. This approach was demonstrated by developing an integrated prototype that dynamically created cultural behavior in a virtual environment.

15. SUBJECT TERMS: cultural modeling, behavior modeling, virtual training environments, behavior moderators, artificial intelligence, cognitive modeling, simulated environments

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON (Name and Telephone Number)
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unclassified		Diane Hadjiosif 703 602-8047

Research Note 2009-02

Culturally Aware Agents for Training Environments (CAATE): Phase I Report

Scott Neal Reilly, Clare Bayley, David Koelle,
Stephen Marotta, and Jonathan Pfautz
Charles River Analytics, Inc.

Michael Keeney,
Aptima, Inc.

Michael J. Singer
U.S. Army Research Institute

ARI-Orlando Research Unit
Stephen L. Goldberg, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926

January 2009

Army Project Number
622785.A790

Personnel Performance
and Training Technology

ACKNOWLEDGMENT

This work was performed under Government contract number W91WAW-07-P-0250. This work was funded in its entirety by the Office of the Secretary of Defense (OSD).

CULTURALLY AWARE AGENTS FOR TRAINING ENVIRONMENTS (CAATE): FINAL REPORT

CONTENTS

CONTENTS (cont'd)

Page

LIST OF FIGURES

Introduction

During the last decade, the U.S. Army has conducted a wide range of missions within the context of very different cultures and languages. Missions have been conducted to enforce peace agreements, to provide humanitarian assistance, and to provide disaster relief. Such missions are often carried out as part of a multi-national force. Key to these efforts is the application of Full Spectrum Operations, in which missions are executed in the context of cultural understanding (Chiarelli & Michaelson, 2005). These missions also often require junior leaders and Soldiers to interact and communicate effectively with people whose cultures, languages, lifestyles, and beliefs are very different from those found in the U.S. It has become increasingly clear that strategy, operational plans, and implementing procedures must occur in the context of local cultural constraints to be effective (Stofft & Guertner, 1995).

To achieve a higher level of mission execution performance, it will be necessary to train Soldiers to understand and work with a sound knowledge of the relevant local cultural context. Performing this kind of training using live role-playing is possible, but does not reasonably scale to the needs of the U.S. Army. Computer-based training, however, does scale if used properly. Unfortunately, computer simulations of culturally situated people are currently very limited in fidelity, making them unsuitable to the task. One reason for this deficiency is that the task of creating fully culturally aware models of humans is extremely difficult and, even if possible, may not be affordable using currently available tools.

We believe that this task is achievable, at least to a degree that is useful for the training needs of today's culturally situated Soldier. By focusing the modeling task on those elements of culture and cultural behavior that *most* impact the successful execution of tactical missions, we can constrain the problem without sacrificing the key training goals.

The planned approach is to focus the modeling effort through a two-step analysis. First, we need to identify the key cultural factors that most influence human behavior in general, and the kinds of behaviors that are likely to influence mission success in particular. These should include factors such as group affiliation, respect/face, or gender role expectations. The second step is to identify the means by which such underlying factors affect behavior and are communicated by behavior. For instance, the current emotional state will often be displayed in culturally unique ways, such as through gaze, body position, gestures, and paralinguistic utterances. If Soldiers are not trained to recognize and communicate using these behaviors, they will likely not be able to understand or effectively respond to, for instance, operationally critical emotional cues (e.g., building anger, distrust) within unfamiliar cultural contexts. We will accomplish these analyses using a mission essential competencies (MEC) approach that was originally developed for understanding and training piloting skills and now is being put to use for much more complex interactions between the pilot and the Air Operations Center (AOC) (Barsch et al., 2007). During Phase I, we demonstrated that this basic approach can also be applied to analyzing social and cultural skills.

While this process of constraining the modeling problem is an important first step, the difficult problem of affordably creating the necessary models remains. This is, again, a two-step process. First, we need to model the underlying factors that affect behavior. Second, we need to

model the behaviors that result from these cultural factors. The first modeling problem will require a rich representation of the social groupings, affiliations, and relationships that affect culturally situated behavior. For instance, we will want to be able to model interconnections such as familial relationships, group membership, and attitudes (e.g., trust, dislike). To accomplish this, our design leverages social network modeling technologies provided by our in-house CONNECT™ tool, which supports the development of rich models of interconnected agents within a graphical development environment. This element of our approach is critical, but often overlooked by more traditional agent-based approaches, such as the Soar-based work out of the Institute for Creative Technologies (ICT) (Johnson & Lester, 2000). The second problem requires a rich human behavior modeling architecture that is capable of modifying behavior based on social and cultural factors. To accomplish this, our design incorporates Charles River Analytics' AgentWorks™ human behavior modeling tool, which provides graphical support for developing (and reusing) simulated agents with behaviors that generate different behaviors based on the cultural context.

During our Phase I effort, we demonstrated the feasibility of this approach. Using our in-house CONNECT social network modeling and reasoning tool, we demonstrated that it is possible to model the rich social and cultural knowledge and relationships that underlie culturally based behavior choices. Also, using our AgentWorks graphical human behavior modeling tool, we demonstrated that it is feasible to model realistic social and cultural behavior. We also created mockups for tools that are based on these existing tools but that are expanded to more directly support affordable modeling of cultural behaviors, which demonstrate the feasibility of constructing tools that enable developers to build cultural agents affordably.

Finally, we designed and prototyped a runtime architecture that supports integrating CAATE-based cultural agents with a variety of simulation environments in a modular manner. This approach that makes it possible to port agents from one environment to another, even when the environments vary considerably in their ability to generate the necessary outputs (e.g., facial expressions) and accept the necessary inputs (e.g., gestures) that are important for cultural training applications. We performed an evaluation of a number of simulation engines based on their ability to provide cultural expressiveness (e.g., gesture, facial expressions) and on their support for integration with third-party agent-control architectures. Based on this analysis, we chose to use the Half Life 2 engine for our Phase I technical demonstration of the feasibility of our approach. For future efforts, however, such as a Phase II follow-on, we recommend using at least one other simulation engine to demonstrate the portability of our design.

In the Background Section, we provide an overview of existing, relevant technologies and related efforts. In the Phase I Technical Objectives Section, we describe the objectives of this Phase I effort. In the Phase I Technical Approach and Results Section, we describe the results of the Phase I effort. In the Conclusion Section, we summarize our results and recommendations.

Background

This section details key background material relevant to our Phase I effort and our Phase II recommendations. In the Characterizing Cultural Differences: Cultural Dimensions Section,

we describe existing work in organizing cultural and social traits and behaviors. In the Cross-Cultural Personality Traits Section, we briefly discuss previous work in representing personality traits in a structured manner. In the Simulation Environments Section, we describe the results of our initial survey of existing simulation environment platforms for use in virtual training applications.

Characterizing Cultural Differences: Cultural Dimensions

A number of disciplines have attempted to construct models of specific cultures in terms of a limited set of factors or characteristics—most notably cultural anthropology and social psychology. These models represent sources of factors that may, or may not, be useful in constructing predictive models of group or individual behaviors. Probably the best known set of factors is that of Geert Hofstede (Hofstede, 1980), who studied 72 different cultures (countries) within a particular organizational setting (IBM). Hofstede identified 5 cultural 'dimensions': power distribution, uncertainty avoidance, individualism vs. collectivism, femininity vs. masculinity, and short vs. long-term time orientation. Another set of factors has been defined by Fons Trompenaars: universalism vs. pluralism, individualism vs. communitarianism, specific vs. diffuse codification of knowledge; neutrality vs. affectivity (referring to acceptability and desirability of displaying emotions); inner vs. outer directed; achieved vs. ascribed status; and sequential vs. synchronic time orientation (Trompenaars, 2001). The anthropologist Edward T. Hall has identified yet another set of factors that distinguish among cultures, and which are particularly relevant during cross-cultural communication: *time orientation, context, and space* (see (Hall, 1977; Hall, 1966) for more detail). As discussed above, research to date has not shown that this level of analysis yields particularly useful descriptions of behavior moderators for incorporation into human behavior models.

Cross-Cultural Personality Traits

The most extensively-studied personality model is the five factor model (Big 5: extraversion, stability, openness, agreeableness, and conscientiousness) (Costa & McCrae, 1992), and these researchers have conducted cross-cultural assessments and have concluded that the Big 5 factors are indeed valid across cultures (McCrae, 2000). McCrae and colleagues also report that traits follow the same patterns of developmental change in adulthood across cultures. A number of studies reveal interesting specific supporting arguments (e.g., Williams, Satterwhite, & Saiz, 1998). While these results appear reasonably robust, some words of caution are warranted, due to the methodologies used to obtain the results. Specifically, Triandis and Suh (2002) caution about the fact that most of the subjects were students, and that educated individuals may not be comparable with less-educated or illiterate individuals. Furthermore, cultural anthropologists have developed alternate, more culturally-oriented, personality inventories (La Rosa & Diaz-Loving, 1991 as cited in Triandis et al., 2002) and found little correlation between these inventories and the Big 5. These traits, given that they represent a different level of analysis than broad cultural dimensions and have clear methods for assessment, may provide a better foundation for the examination of cultural influences on behavior.

Simulation Environments

While the focus of our effort was not on simulation engines themselves, it was important for us to understand the current state of the art, both so we could design our solutions appropriately and so we could choose a suitable platform for a demonstration prototype. To this end, we compiled a list of current simulation environments, designed a simple set of evaluation criteria, and evaluated a number of promising candidates. Appended to this report is a list of current simulation environments. Due to the priorities and resource constraints of this effort we were not able to complete our analysis, though this table does provide an initial set of evaluation criteria and some information with respect to the listed engines as evaluated against these criteria. In addition, appended to this report is a more thorough analysis of the four game engines that we evaluated for our purposes that seemed most feasible for use as our demonstration platform. Based on this analysis, we chose to use the Half-Life 2 engine for the Phase I demonstration as described in the Design and Demonstration of Modular CAATE Runtime Architecture Section.

Phase I Technical Objectives

The primary objective of the Phase I work was to design and demonstrate the feasibility of Culturally Aware Agents for Training Environments (CAATE), a tool for the affordable development and deployment of culturally believable agents for simulation-based training applications. Specific objectives included answering the following questions:

- **Training Objectives and Methodology Identification:** What are the key challenges for training culturally situated mission behaviors? What skills and knowledge need to be taught in order to maximize effective mission execution? What skills and knowledge are not important? What are the most appropriate methods for training mission-critical skills and knowledge?
- **Scenario Development:** How should we develop a demonstration scenario based around key difficulties encountered by Soldiers during culturally sensitive interactions? Which broad operation environment would best demonstrate the capabilities of our system, while providing adequate complexity to capture key domain features and remaining amenable to a Phase I effort?
- **Cultural Dimensions and Behavior Identification:** What critical cultural dimensions should be addressed by a training system? What types of cultural dimensions most clearly drive specific human behaviors? How should these factors be represented in a model for generating culturally based behavior? What other factors (e.g., situational information) need to be captured? How do we map from the underlying cultural factors to external behavior?
- **Modeling Methods:** What modeling methods are appropriate for representing human behavior and the impacts of culture on behavior? How can the modeling methods ensure presentation of culturally based behaviors in a manner consistent with the training methodology? How should the model support interaction among cultural factors and situational influences? What methods support validation of the underlying cultural dimensions?

- **Training System Approach:** Does the combination of culturally based behavior models and a simulation-based training environment represent a feasible and cost-effective means of training Soldiers' cultural awareness? How should we combine existing research, in-house tools, and commercial off-the-shelf or Government off-the-shelf (COTS/GOTS) packages to develop a comprehensive training environment? How can the proposed tool be integrated effectively with existing Department of Defense (DoD) training systems and/or simulation environments?
- **Validation and Verification:** How should we validate the impact of the cultural dimensions on specific behaviors? How do we verify that the system is producing the desired behaviors in the correct situations? How do we validate the generation of behaviors for training purposes against the likelihood of those behaviors in the real-world? How do we ensure positive transfer of training?

By successfully addressing these objectives, we demonstrated the feasibility of our approach and recommend future effort focus on the advanced development and thorough evaluation of the CAATE approach and tools.

Phase I Technical Approach and Results

The first task was to design a demonstration scenario**.** We worked with Aptima, the Sponsor, and a cultural subject matter expert (SME), who lived in Iraq for 30 years and has experience as a translator for US and British forces, to identify and design a suitable culturally situated training scenario that includes a set of mission-required interactions with cultural elements. We settled on a "first ten seconds" scenario where the Soldier is responsible for engaging in the initial greeting behaviors that would precede any of a number of operational interactions (e.g., requesting papers at a checkpoint, searching a house in a cordon-and-search operation). The design of this scenario demonstrated our ability to create virtual training scenarios; it also provides us a reusable scenario that can be leveraged in follow-on efforts.

The second task in the approach was to identify critical cultural dimensions and behaviors**.** We worked with our Iraqi cultural SME and existing literature to identify an initial set of cultural dimensions and behaviors. This task served two purposes. First, we performed a general review of the literature to identify a wide range of cultural dimensions and behaviors that our modeling and simulation tools might need to represent and simulate. The initial set of potentially important cultural *dimensions* that we identified, includes: culture, emotions, attitudes, relationships, personality, personal traits, state, social roles, physical context. The initial set of potentially important cultural *behaviors* that we identified includes: gestures, social actions, facial expressions, eye gaze, posture, proxemics, other actions, time-based actions, body state, language, (including semantic choices, lexical choices, syntactic choices, volume, intonation/tone, and accent), use of humor, speech irregularities, time-based language, appearance (including age, gender, and dress), and combinations of these behaviors. Second, we performed a MEC analysis of the scenario developed under Task 1 in conjunction with our Iraqi cultural SME to demonstrate the feasibility of evaluating scenarios for operationally relevant cultural dimensions and behaviors to be modeled. We recommend that future efforts in the area focus on further extending these lists, refining these categories for particular cultural groups, and focusing these lists for successful training in particular operational tasks.

The third task identified for the Phase I approach was to identify and develop a training methodology**.** We investigated and documented methodologies for training and evaluating the types of skills that will need to be trained by CAATE-based systems. As much as possible, these training objectives and the training methodology were designed to be sufficiently general to apply to a range of similar training scenarios. First, we looked at how simulation-based training with culturally realistic agents can be incorporated effectively into a comprehensive training regimen. For instance, virtual environments (VEs) provide a great deal of richness and realism that can, for instance, effectively provide practice for skills, such as culturally-based emotion-recognition skills, that are initially taught in a non-immersive manner. Second, we looked at mechanisms for evaluating the success of CAATE-based training applications, including mechanisms to track the Soldier's performance in the virtual training environment. The goal of this overall process, was to help us understand the requirements for tools for developing such agents and evaluating their effectiveness. We recommend future work in this area should (1) continue to refine and extend this methodology and (2) incorporate our results into our development and implementation efforts which need to be informed by these training and evaluation goals.

In the Phase I effort the fourth task was to design a modeling tool for cultural dimensions and the fifth task was to design a modeling tool for cultural behaviors**.** We designed a modeling architecture and associated modeling tools to support the cultural-dimension and cultural-behavior modeling requirements established under Tasks 1-3. While we initially broke the two modeling requirements (cultural dimensions and cultural behaviors) into two tasks, we came to believe that the modeling process needs to seamlessly integrate both aspects of modeling into a single tool, so we discuss these two tasks together. Based on our experience in this area and discussions with the Sponsor, we identified a number of key requirements for this set of tools, which are presented in detail in Requirements Analysis section, below. The basic requirements for modeling tools are that they be *modular, easily distributed, flexible, easy, affordable, and expressive.* We developed tool mockups based on our in-house CONNECT social network modeling tool and our AgentWorks agent development environment that support these goals. This design provides the basis from which a Phase II effort to develop a full-scope CAATE prototype can begin.

The sixth task was designed to provide a demonstration of the CAATE training system. We evaluated existing simulation environments with respect to their ability to represent the culture-based behaviors defined under Task 1. We selected Half-Life 2 as the simulation environment for demonstration of the culture-based behaviors generated with the CAATE system based on its support for third-party agent controllers and the physical expressiveness of the agents. We demonstrated the technical feasibility of using the CAATE system to generate culture-based behaviors for agents in this environment. The CAATE agents sense their environment and the culturally relevant behaviors being performed by the trainee's avatar, they react to these stimuli, they act in a manner that is dynamically chosen to be appropriate for the situation and their cultural background, and they update the simulation interface to enable the trainee to react to these behaviors. In support of the *modularity* goal described above, we designed and prototyped a mechanism for providing rich cultural and social interactions on a wide variety of simulation platforms, even those with minimal capability of accepting rich user inputs or generating complex visual or audio behaviors. For instance, in simulation environments

where communication isn't feasible (e.g., the agent wants to make a facial expression but the simulated environment doesn't support facial animations), the CAATE Trainee Input/Output window fills in such holes by providing a simulation-independent mechanism for describing these extra-simulation elements that are important for the cultural training experience. This effort helped us reduce the technical risks associated with our design.

In task 7 we were required to develop a plan for validation and verification. We evaluated two kinds of evaluation metrics and evaluation plans to test the effectiveness of CAATE. First, we explored direct evaluation of the tools themselves, focusing on technical properties and features. Second, we explored means of evaluating the tools through the more indirect, but also more important, mechanism of evaluating the training effectiveness of applications built with these tools. We identified a number of general evaluation approaches, such as evaluations with training experts, evaluations with cultural experts, and evaluations with a population of students/trainees. In conjunction with the latter, we also investigated appropriate methods for evaluating the success of training particular kinds of social and cultural skills and how these evaluations could be done effectively within a virtual training application through minimally intrusive trainee tasks that support the evaluation process.

Requirements Analysis

Through discussions with the Sponsor, we identified a set of requirements that any approach to modeling agents for simulation-based cultural training systems should have. The requirements include:

- It should be *modular*. In particular, our solution should be able to integrate with any of a number of possible simulation platforms for virtual cultural training. One by-product of this is that we determined through discussions with the Sponsor that we should not target the One Semi-Automated Forces objective system (OOS) as indicated by the initial solicitation, but should also ensure that our runtime agent models are consistent with, for instance, commercial game platforms such as RealWorld, OLIVE, DARWARS, and Delta3D.
- The resulting agent-based training applications should be *easily distributed*. For training applications developed with these tools to be useful to the U.S. Army, they will need to be widely distributed. Our solution should place as few costs and other constraints on the distribution of the runtime models as possible. While we were able to make significant progress during this effort through leveraging existing Charles River Analytics' intellectual property, we have also made an effort to ensure that the resulting training applications can be widely distributed by the U.S. Army.
- It should be *flexible*. The tools we provide must be sufficiently flexible to enable the development of new content, including new training scenarios for new interactions and in new social and cultural dimensions and behaviors as needed to support these new training applications.
- It should be *easy and affordable* to use. Instead of forcing new content to be built for each new training scenario, we will provide methods for reusing content as much as possible, including reuse of cultural dimensions, behaviors, and agent profiles. In

addition, the tool itself should provide graphical user interfaces where feasible to enable simpler construction of cultural agents.

- It should be *expressive*. It needs to be able to model a wide range of cultural and social phenomena. Two issues that we identified as being important and often overlooked in social-agent modeling systems are (1) the agents need to be able to have multiple simultaneous behavior moderators (e.g., an agent can be an elderly fundamentalist Sunni Iraqi woman), and (2) the agents need to be able to express directed behavior moderators (e.g., relationships, emotion towards a particular person). Most existing behavior modeling systems that support behavior moderators do not realistically model multiple moderators (Neal Reilly et al., 2007) and model operators as a type and value (e.g., stress of 75%) but without the ability to direct the moderator, which can impact, for instance, who the corresponding behavior is directed towards (Neal Reilly, 2006).

General Process Architecture

Figure 1 shows the overall training-application development process that we envision being used for simulation-based training applications for cultural and social skills. The process begins with the relevant training objectives in terms of missions, scenarios, and operational skills. We performed a *MEC Analysis* to derive the relevant *Soldier Knowledge, Skills, and Abilities (KSAs)* to be trained, including recognition skills (e.g., recognizing anger by evaluating the gestures and speech patterns of others) and active behavior skills (e.g., using the proper form of address to show respect). Other important results of this process are effective means of training the key skills as well as evaluation metrics for those skills.

The recognition skills provide requirements for the *NPC (non-player character) Behaviors* and *Cultural Factors* that are modeled by the computer. We identify the underlying *Behavior Moderators* (e.g., cultural factors) that give rise to important behaviors as well as the behaviors themselves. The behaviors and cultural factors are modeled using the *Behavior & Moderator Modeling Tools*, which are built on Charles River Analytics' CONNECT (for social relationship modeling) and AgentWorks (for human behavior modeling, including moderated behavior) tools. The resulting *Runtime Behavior Models* are used to drive the runtime agents that the Soldier interacts with in the virtual setting.

The active skills provide requirements for the interface to the system. For instance, if we want to train Soldiers in using the proper form of address in a speech greeting, we need an interface that supports spoken inputs. These requirements help to determine which of a number of simulation platforms are most appropriate. A *Simulation Platform Choice* is made and a virtual environment is created (or reused from a previous effort) for the specific training application. In reality, however, the range of interface needs, both the

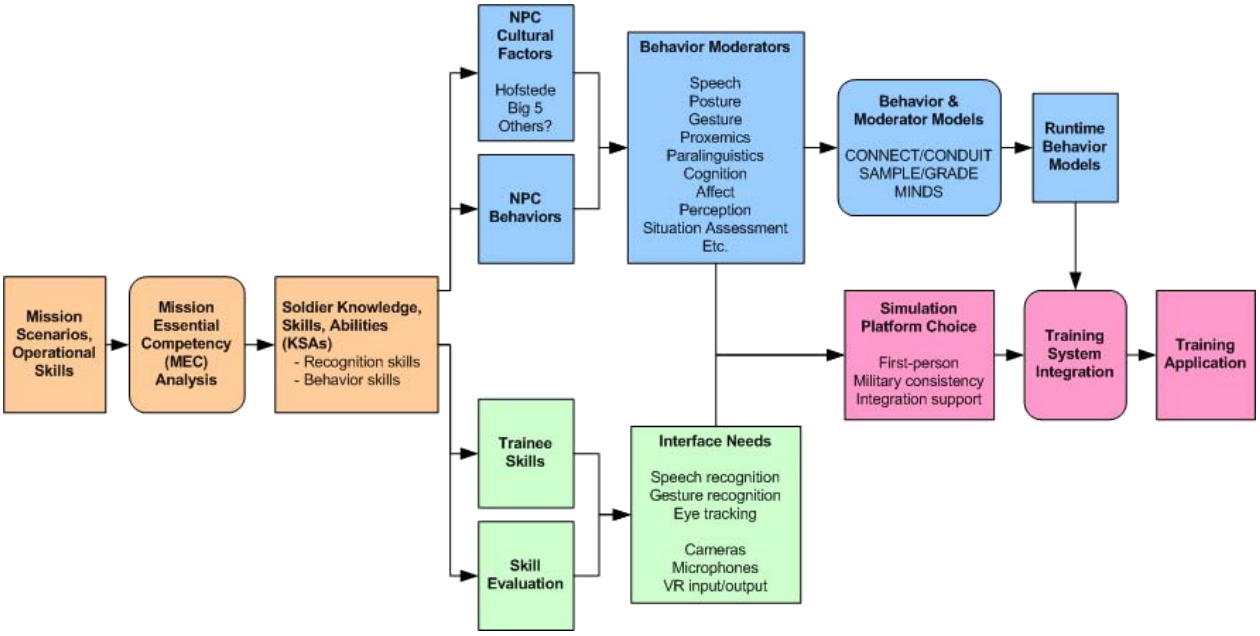


Figure 1. CAATE cultural training system development process.

required trainee inputs and the required agent outputs, are rarely fully met by the chosen simulated environment. To this end, the *Training System Integration* process provides additional supports for these extra-environment inputs and outputs as we describe in the Design and Demonstration of Modular CAATE Runtime Architecture Section. The simulation system interacts with the trainee and the simulated agents at runtime to provide the overall training experience.

The CAATE effort is focused on the MEC process and the development tools for modeling and simulating culturally realistic agents. We are also concerned with ensuring that the agents that we develop are portable to a variety of simulation environments, though, at the Sponsor's direction, this effort is less focused on the choice of or the creation of such environments and interfaces.

During Phase I, we developed a simple training scenario to help ground our designs and discussions. To this end, we investigated an interaction used to gather information during the initial phase of an encounter with a member of a local population. Some of the advantages of such a scenario are that it is relatively simple and almost scripted in its regularity, it is used frequently and in many different contexts (e.g., at a checkpoint, during a dwelling search), and it provides an opportunity to investigate a number of interesting differences in behavior based on social and cultural differences (e.g., interaction with a male vs. female, old vs. young, in their house vs. in the street). Dr. Keeney at Aptima organized and conducted initial MEC analysis interviews with the cultural SME who lived in Iraq for 30 years and has experience as a translator for U.S. and British forces.

Review of Critical Cultural Dimensions and Behaviors

During Phase I, we extended our team's expertise with a literature search and an interview with a cultural SME to develop an initial set of cultural and social dimensions that can affect behavior, and an initial set of behaviors that display cultural and social based variation. During Phase I, we made an intentionally broad search. Part of our recommendation for a Phase II effort is to prioritize these sets based on their impact on operational success. In the Process Section, we summarize the process we used to generate these lists. In the Results Section, we summarize the results of our analysis and search.

Process.

Review of person-perception research, especially for culturally-based differences. We reviewed the social cognitive and person-perception literature (e.g., Wyer & Srull, 1994). We sought empirically-based information about how humans perceive the emotions, actions, and intentions of other people, and in particular, information about variables that a model of cross-cultural behavior would need to represent. A finding from this review inspired confidence that previous research findings of cultural differences in perceiving emotions in others may be only methodological artifacts (Elfenbein & Ambady, 2002a, 2002b, 2003a, 2003b, 2003c), which suggests that techniques used to create computer-based images of human facial expressions representing individuals in Western cultures should be capable of producing realistic images of members of non-Western cultures. We noted that while behavior in others is visible, the motivation behind it is not (Gilbert, 1995). Consequently, humans are likely to infer situational (external) reasons for their own behavior, while attributing behaviors of others to stable, internal causes, in processes that are fast, nonlinear, and largely operate automatically and outside conscious thought (Lewicki, Hill, & Czyzewska, 1992; Logan, 2005; Moors & De Houwer, 2006; Vallacher & Nowak, 1994, 1997). This suggests that it is important to include situational factors in the model, not only those relating to internal characteristics such as personality or attitudes. Woods, Pease, Stout, and Lacey (2006) offer a large-scale illustration of the implications of effects of situation, in which illuminating the situation from the perspective of Saddam's senior leaders brings a degree of rationality to their pre-invasion planning. We also noted that emotions and affect will moderate processing of information (Clore, Schwartz, & Conway, 1994; Ortony, Clore, & Collins, 1988; Ekman, 1992; Ekman & Davidson, 1994). Thus, computer models must include not only stable personality and cultural factors, but also the variable and transient effects of emotions. We noted during our review of requirements for software validation that authors had recognized the complexity of human behavior (Goerger, McGinnis, & Darken, 2005; Pew & Mavor, 1997, and the special topic section on validation of human behavior representation in DMSO, 2006). These sources note that in contrast to models of physical objects and processes, human behavior is nonlinear, interactive, and contains elements of randomness, which psychologists have recognized for some time (Guastello, 1995; Whitehead, 1938).

Review of research in non-verbal communications, especially for culturally-based differences. We next reviewed the literature on non-verbal communications. We noted from this review that non-verbal behavior can repeat, conflict, compliment, accentuate and moderate, and substitute for verbal expression (Knapp & Hall, 1996). We noted that non-verbal behavior can

include not only eye movement, facial expressions, gestures, and posture, but also variables external to an individual such as clothing, layout of physical environments, and markings of territory and personal space. As noted earlier, inferences in non-verbal communications can also operate largely outside conscious thought.

Review of open-source information on Iraqi and Arabic-Islamic culture and history. We next review references that specifically focused on Iraqi and Arabic-Islamic history and culture. Our goal was to be reasonably comprehensive while focusing on sources likely to be familiar to warfighters in an effort to confirm or refute this information. We sought to include information only when clearly supported by multiple sources. Examples of sources and works we reviewed included (Al-Shawi, 2006; Gilsenan, 2006; Kepel, 2004; Patai, 2002; Polk, 2005; Pryce-Jones, 2002).

Review of historical sources describing warfighter behaviors that can stimulate or aggravate insurgency. The ongoing insurgency in Iraq is not without historical precedent. We reviewed several works that suggest that past experiences can offer suggestions for successfully reducing the degree of opposition to an American presence in another country. Among these references were (Al-Shawi, 2006; Boot, 2002; Boyle, 1994; Hashim, 2006; McPherson, 1988; Mackey, 2004; Poole, 2005). Two findings from this review were noteworthy in highlighting our apparent lack of learning from history. Many of the same behaviors exhibited by American warfighters in Iraq have produced the similar outcomes even without large cultural differences. The British experience in Mesopotamia offers remarkable similarity in goals, methods, and outcomes, yet references describing Iraqi history provided to our forces typically provide only fleeting mention of this experience. Thus, for Americans, events in Iraq became a first-time experience, but in the Iraqi memory, all this has happened before.

Consideration of competencies from related jobs. Our review of lessons learned from past insurgencies suggests that warfighters need to supplement their skills by adding several typically required of police officers. Consequently, we reviewed the entry for police officer in the O*Net database (http://online.onetcenter.org/), which, as our primary source of occupational information, provides comprehensive information on key attributes and characteristics of workers and occupations. We identified 31 potential areas of expertise that could supplement traditional military skills, such as how to communicate with persons outside an organization such as members of the public, community relations, dispute resolution, identifying suspects, negotiation, and protecting people and property.

Confirmation of information from multiple sources and subject-matter experts. We confirmed any recommendations for variables and characteristics to be included in the simulations by two means. First, we included only data provided from at least two independent sources. Second, we asked an expert on Iraqi culture and two experts on U.S. Army culture (specifically, former warfighters with experience in Iraq) to review the final list of variables and characteristics. Thus, we vetted the proposed cultural variables and characteristics using at least three sources.

Results.

The set of cultural and social dimensions we identified during Phase I is summarized as follows:

- **Culture.** Many interesting cultural variations can be captured with a set of dimensions, such as Hofstede's individualism, power distance, uncertainty avoidance, masculinity, and long-term orientation (Hofstede, 1980), Hall's high/low context and monochronic/polychronic (Hall, 1977), or Trompenaars and Hampden-Turner's universalism versus particularism, communitarianism versus individualism, neutral versus emotional, diffuse versus specific cultures, achievement versus ascription, and human-time relationship versus human-nature relationship (Trompenaars, 2001; Hampden-Turner & Trompenaars, 1997). Such broad strokes do not, however, specify many important aspects of cultural variation, including many customs, so it will also be important to model more specific cultural traits as well, such as ethnicity, national origin, and religion.
- **Emotions.** It may be possible to use a set of common (or basic) emotions, such as anger, disgust, fear, happiness, sadness, surprise (which combines Ekman's (1992) and Oatley & Johnson-Laird's (1987) sets of basic emotions—there are more than 10 different sets that have been postulated in the literature). We will also need to include non-basic emotions (e.g., jealousy, pride, shame, gratitude) and more culturally specific emotions that can be important from a training perspective (e.g., Japanese amae or Malaysian amok). The direction (e.g., who the NPC is angry at) and the cause (e.g., why the NPC is angry) can also influence behavioral choices and, therefore, will likely need to be represented.
- **Attitudes.** Attitudes include, for instance, anti-Americanism; there is no universal set of attitude types, so these will need to be added as they are found to be needed for training scenarios.
- **Relationships.** We will want to model, at least, familial, organizational, and social relationships.
- **Personality.** As with culture and emotion, there are dimension-based theories of personality that can capture many important aspects of personality for training simulations, such as the Big Five personality traits (Neuroticism, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience) (Goldberg, 1993; Costa et al., 1992) which can be expanded to include other traits from the literature (e.g., religiosity, manipulativeness, honesty, thriftiness, conservativeness, masculinity, snobbishness, sense of humor, identity, self-concept, and motivation) as needed. Though, again as with culture and emotion, there are many aspects of personality that are not determined by these traits and will need to be added if required by particular training scenarios.
- **Personal traits.** We will want to model, at least, age, gender, and dress.
- **State.** This is any temporary feature of the agent that might affect the agent's behavior or the way that the Soldier should interact with that agent (if observable). For instance, hunger, fatigue, illness, being hurt, or being in a hurry or busy all fall under this category.
- **Social Roles.** This is any role that the agent plays in the society. It will be common for agents to play more than one social role and the effect of this dimension is typically

mediated by the current context. For instance, a police officer will often act differently because they are a police officer, but are more likely to do so when they are in uniform and on duty. The same agent might also play the role of a father in some contexts, especially when his children are present.

- **Physical Context.** Where a behavior is being performed can affect the performance. For instance, being in a mosque vs. on the street vs. in one's home can affect behavior.

The set of behaviors that can demonstrate social and cultural variability are:

- **Gestures.** Another type of action, primarily for social/communicative purposes. For instance, grouping the thumb and fingers together and shaking it up and down, fingers pointing upwards, indicates "wait" in many Middle Eastern countries.
- **Social Actions.** These are actions that physically involve another person. For instance, in Arab countries it is common for people of the same gender to hold hands while walking as a display of friendship. In the Middle East, Asia, Africa, Italy, and South America, touching others is common. In Japan, the United States, the United Kingdom, and Australia, touching is less common. Other countries, such as France, China, and India, fall in the middle.
- **Facial Expressions.** Many emotions are expressed through facial expressions. Some are universal (see Ekman, 1992), others are culturally specific.
- **Eye Gaze.** In American culture, direct eye contact during a conversation shows respect and intensive listening; avoiding it is a sign of nervousness or lying. But for Koreans, especially older Koreans, direct eye contact is interpreted as aggressive and rude. In many cultures it is respectful to not look the dominant person in the eye, but in Western culture this can be interpreted as being "shifty-eyed," and the person is judged badly because "he wouldn't look me in the eye."
- **Posture.** Displaying the sole of one's foot by sitting with ones feet up is considered rude in many Asian cultures. Standing with crossed arms indicates opposition or defensiveness in many cultures in stressful situations (it can also mean the person is cold or is giving an idea serious thought in other contexts).
- **Proxemics.** The distance between people in a conversation differs from culture to culture. People in Middle Eastern cultures tend to stand more closely together than is comfortable for Westerners, though moving away can appear rude.
- **Other Physical Actions.** These are physical behaviors that don't fall under other categories (e.g., gestures, posture). For instance, an NPC might walk away in the middle of a conversation. Also, an NPC might offer a drink or a snack to teach acceptance of hospitality. Saying "No thank you" in response, even if said very politely, can be offensive. Instead, accepting the offering even if you choose not to eat or drink it is considered more acceptable than rejecting their hospitality.
- **Time-Based Actions.** The temporal aspects of actions, such as being punctual/late, acting quickly/slowly, or getting to the point quickly/slowly are important to recognize. These are not actions themselves, but affect how other actions are carried out.
- **Body State.** Various emotions are expressed through other changes to the body state, such as sweating, twitching, changes in breathing rate, muscle tensing, leg shaking, and coloring (e.g., turning pale/red).
- **Language.** Many features will be expressed through aspects of language, including:

- **Semantic choices.** What the NPC chooses to say or not say is extremely important for mission success and will be affected by their culture, personality, mood, and other factors.
- **Lexical choices.** Choice of words, such as crude words or uncommon words, can indicate a range of factors, from social class to mood.
- **Syntactic choices.** Complex vs. simple syntax and syntactic errors when speaking can indicate level of education and level of liking/cooperation (uncooperative NPCs will typically use short responses).
- **Volume.** In regards to vocal emphasis and volume, people in the Middle East may communicate in ways which Westerners reserve for when they are angry or upset.
- **Intonation/Tone.** NPCs can use sarcasm or emotional intonation to indicate dislike, attitudes, and emotions.
- **Use of humor.** Jokes and humor are very culturally dependent and can be a source of misunderstandings.
- **Speech irregularities.** NPCs might stutter either due to nervousness or personal speech dysfluency, or might use "uh," "um," or a culturally appropriate variant when nervous.
- **Time-Based language.** As with "time-based action," this is not a separate type of behavior, but an important moderator on how linguistic communication happens. Speaking slowly or quickly differs based on culture and emotional state. Also, the willingness to interrupt is based partly on culture. For instance, in Western cultures interruption is often seen as an enthusiastic participation in the conversation (especially between males); in Eastern cultures it can be considered rude and it is sometimes preferable to pause before answering to show that you are considering the question/point.

- **Combinations.** It is important to be able to express multiple combinations at once. Not only does this provide additional realism to the simulation, but a key skill to learn will be to interpret ambiguous actions in context. For instance, folded arms can suggest defensiveness or simply being cold and other environmental and behavioral cues need to be evaluated to properly interpret this pose.
- **Perception and interpretation of the above.** The NPCs need to be able to both generate these behaviors themselves to indicate their cultural/social background and to recognize them when they are performed (or miss-performed by the Soldier). While some of these behaviors will be hard to generate on the part of the Soldier (e.g., many aspects of the body state are not consciously controlled), we include them anyway for completeness in case they can be modeled in the future (e.g., using physiological sensors on the Soldier to measure stress and anxiety).

Design of CAATE Development Environment

During Phase I, we designed and demonstrated the feasibility of an approach to modeling and simulating culturally realistic agents based on existing COTS tools. Figure 2 presents our current design.

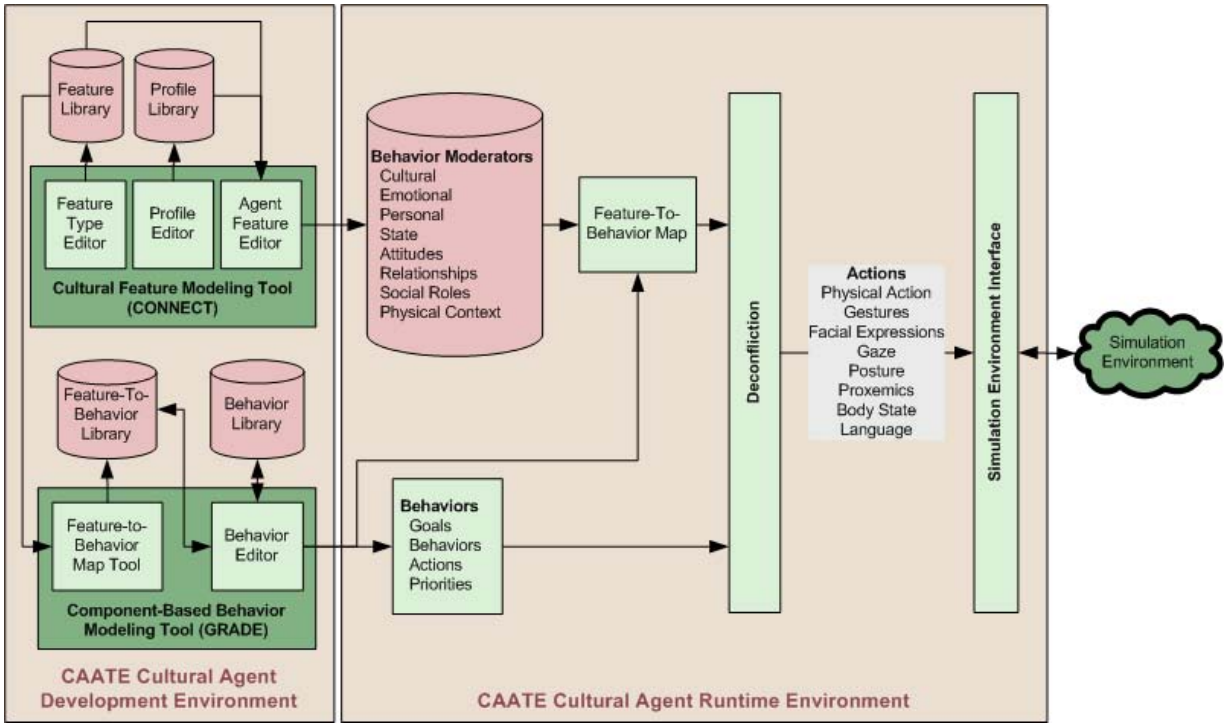


Figure 2. CAATE development and runtime architecture.

The CAATE Development Environment (on the left) is built on top of two existing Charles River Analytics tools: CONNECT, a network modeling tool, and AgentWorks, an agent/behavior modeling tool. The typical simulation developer will begin with the *Agent Feature Editor*, which is used to create instances of simulated agents. These agents have individual profiles and a variety of relationships with the other agents and the trainee. The profiles and relationships will largely be chosen from the *Feature Library* and the *Profile Library* and then customized to the particular training scenario. The former includes definitions of standard social and cultural features (e.g., male/female, Arabic, outgoing); the latter includes typical agent types that are reused in multiple training applications (e.g., a typical uncooperative Iraqi Sunni merchant). When existing features or profiles do not exist, the developer can create them using the *Feature Type Editor* and the *Profile Editor*, which share a common user interface with the *Agent Feature Editor*. The ultimate result of this step is a set of *Behavior Moderators* for each simulated agent that affects behavior at runtime.

Once the agents, their social and cultural contexts, and their personal profiles have been defined, the agent developer will create the behaviors that control the agent in the training scenario. The primary tool for defining these behaviors is the *Behavior Editor*, which will be used by the training application developer. There are two types of behaviors that are defined by this tool. The first are scenario-specific *Behaviors* used by the agent. The second type provides the *Feature-to-Behavior Map*. These behaviors are driven directly from the current state of the agent's *Behavior Moderators*. For instance, there will likely be behaviors to map from different emotional states to prototypical facial expressions. This mapping also needs to account for the simultaneous existence of multiple competitive or reinforcing behavior moderators using

methods such as those described in (Neal Reilly et al., 2007). Both types of behaviors are created from within a common user interface. Both types of behaviors can be stored and reused in subsequent training applications. The scenario behaviors are stored in the *Behavior Library* and the feature-to-behavior map behaviors are stored in the *Feature-To-Behavior Library*. We expect that the simulation developer will typically be able to build the mapping behaviors directly from the library. Where this is not the case, they will use the *Feature-To-Behavior Map Tool* to create new mapping behaviors or to modify existing behaviors.

This approach provides two primary sources of behavior at runtime: scenario-specific behaviors and cultural/social default behaviors. The scenario-specific behaviors are also designed to respond to *Behavior Moderators* and display culturally and socially appropriate behavior. These two sources of behavior are passed through a *Deconfliction* module that will typically allow the scenario-specific behaviors to preempt the default behaviors where there are conflicts. For instance, if an agent is in the angry state it might cross its arms based upon a *Feature-to-Behavior Map*. If the agent simultaneously chooses to point as part of the ongoing scenario-based social interaction, the pointing action will occur instead of the (inconsistent) cross-arms action.

The result of this process is a wide variety of different types of action and behavior that the agent can take. These are passed to an interface module that handles the integration with the particular simulation environment. If at all possible, we do not want the agent design to be dependent on the simulation environment being used. This will enable us to more easily switch from one simulation environment to another. There needs to be some sort of simulation-environment-specific integration, which is provided by the *Simulation Environment Interface*. This interface will be built once per simulation environment type that is to be used (e.g., once for OneSAF, once for OLIVE, once for Half-Life).

During Phase I, we also developed mockups of the tools used as part of this development process. Two of the primary goals of these tools are ease of use and the ability to reuse cultural models and behaviors. Figure 3 shows the construction of a simple training scenario. In this case, the training application developer has dragged two computer-controlled agents, Sahib and Raja, from the library on the left onto the workspace. These agents have been previously constructed with a variety of standard social and cultural features pre-set. If none of the existing agent profiles are appropriate for the current training application, the developer can create and save new profiles or modify one of the existing profiles. The developer also uses a tool based on our in-house social-network modeling tool, CONNECT, to identify the "links" between the agents and each other and the trainee. In this case, we have specified a relationship link ("Parent of"), an attitudes link ("Respectful to"), and two emotion links ("Fearful of" and "Angry at").

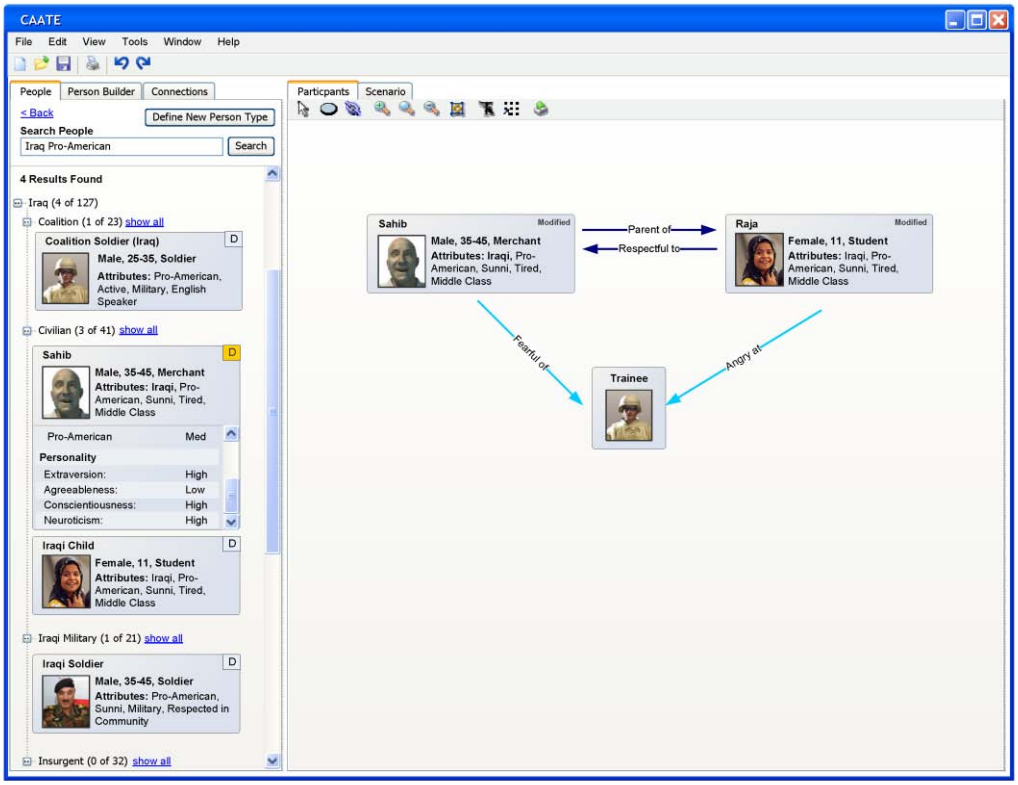


Figure 3. Creating a scenario from existing profiles.

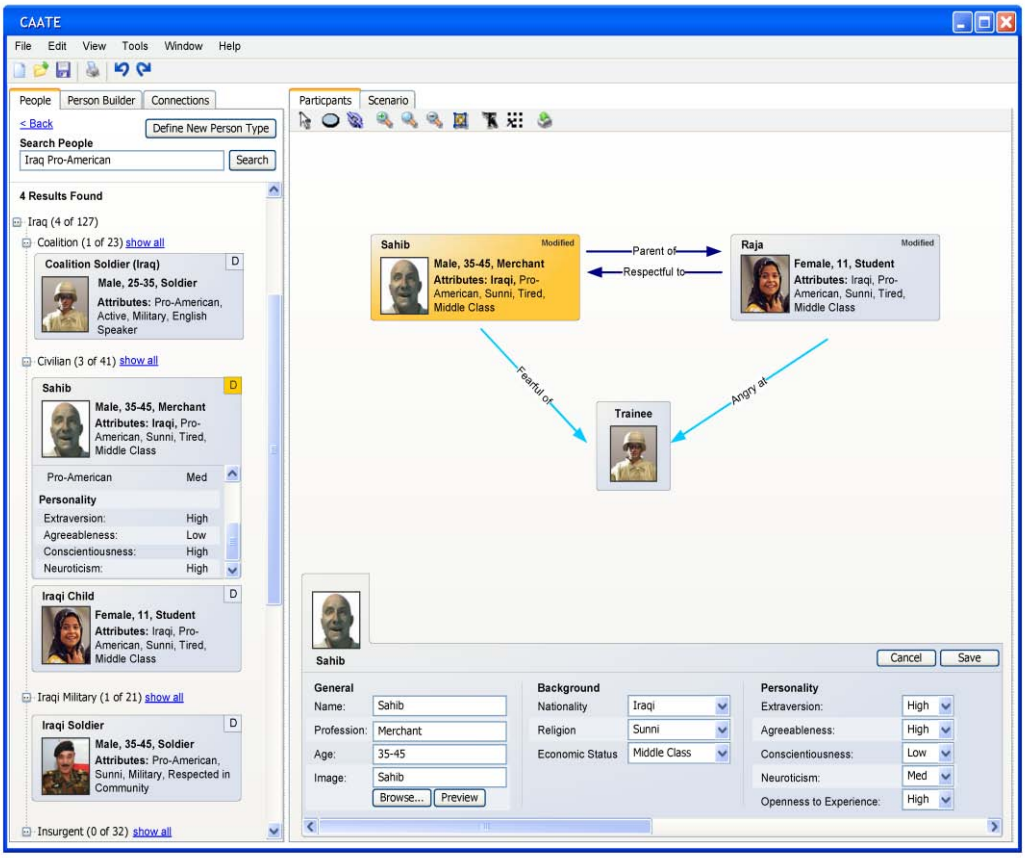


Figure 4. Modifying a profile.

In Figure 4, the developer is modifying the traits of the Sahib agent to reflect the cultural, social, and personal traits desired for the current training application. Once done, the modified Sahib can be saved as another profile for use in future applications as well. The traits will influence the way that Sahib acts and treats others in the environment as we will discuss below.

In Figure 5, the developer is modifying one of the links between the agents. In this case, the intensity of the "Angry at" link is being modified to make Raja initially "Very Angry" at the trainee. Links are also structured entities and can have any number of relevant features. In this case, we are displaying the "Level" or intensity, but emotional links might also include information about, for instance, why the emotion occurred or how quickly it should decay. The types of links and their properties also influence how the agents will behave in the training simulation.

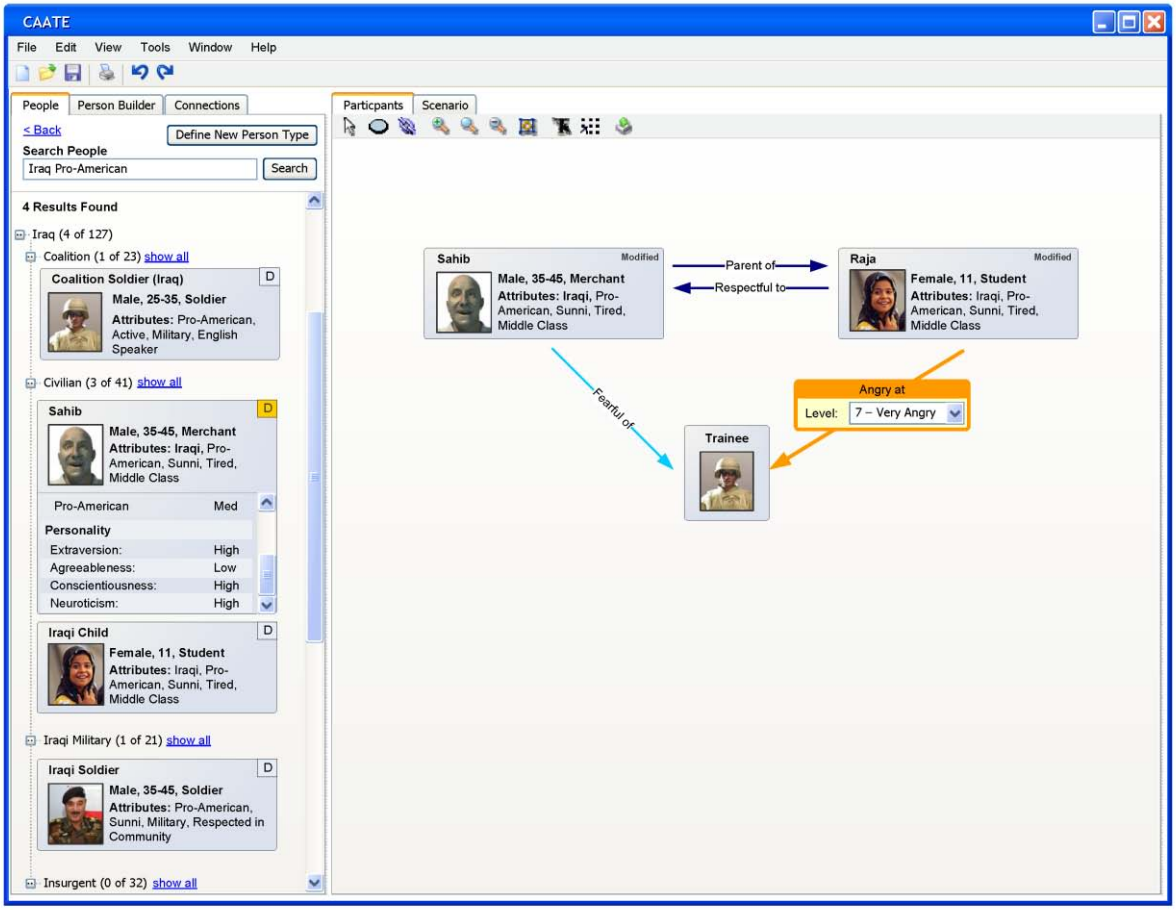


Figure 5. Modifying a link.

In Figure 6, the developer has opened the "Connections" tab on the left-hand side of the application window. This tab enables the developer to control which links are currently visible. This is important for managing scenario development where there are many links of different types. This tab also enables the developer to create new link types. The new type will then be available for modifying the agents' behaviors, which we turn to in the next section.

Figure 7 shows a mockup of the behavior development component of CAATE where the developer is creating and editing the behaviors for the two agents whose relationships were just defined. The developer is able to choose behaviors from a pre-defined library of culturally aware social behaviors. The behaviors are dragged and dropped onto the relevant agent. The developer is able to create reusable behaviors using two possible approaches. First, behaviors can be built to include all possible cultural and social contexts. Such a behavior could be dragged and dropped into a wide range of possible agents (e.g., an elderly Arab male, a young German girl) and it will have the ability to adapt accordingly. Second, and we believe likely to be more common in practice,

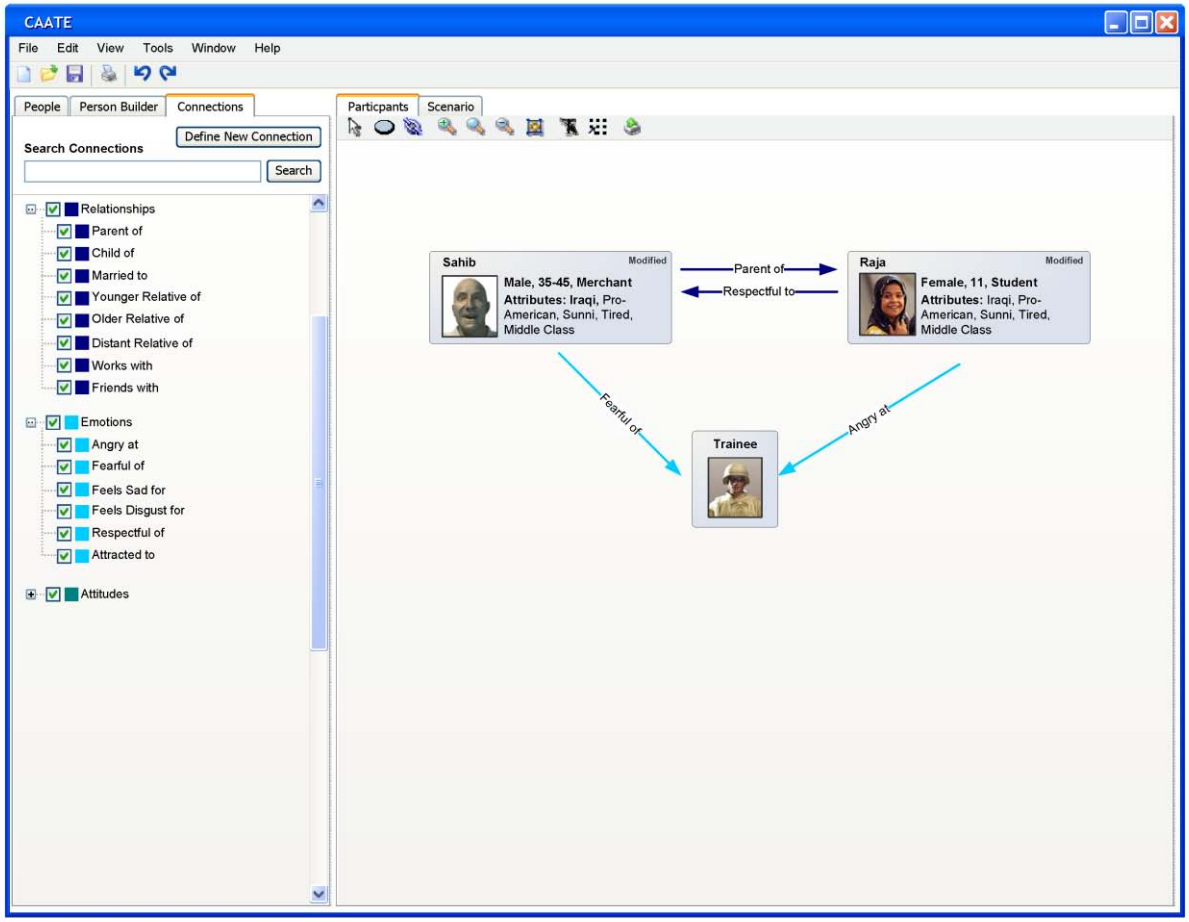


Figure 6. Browsing links.

is that general-purpose behaviors will be built to act believably in a wide range of social and cultural contexts based on high-level abstractions of those contexts (e.g., culture is described in terms of Hofstede's cultural dimensions or personality is described by the Big 5 personality types). More specific behaviors will then be built to suit more specific cases (e.g., Arabic cultures or even Iraqi-specific behaviors as needed). These can be combined with the general purpose behaviors by the developer. For instance, when creating the "first ten seconds" scenario, the developer might use a "Universal" greeting behavior to manage proxemics, which is how close the agents stand when interacting based on high-level cultural abstractions. Such a behavior, however, need not be fleshed out to include all of the linguistic variations used during greetings in specific cultures. An Arab or Iraqi behavior would manage those aspects. As a new culture was to be modeled, the universal behavior could be reused, but a new country or region specific behavior would be created for a non-Arabic agent.

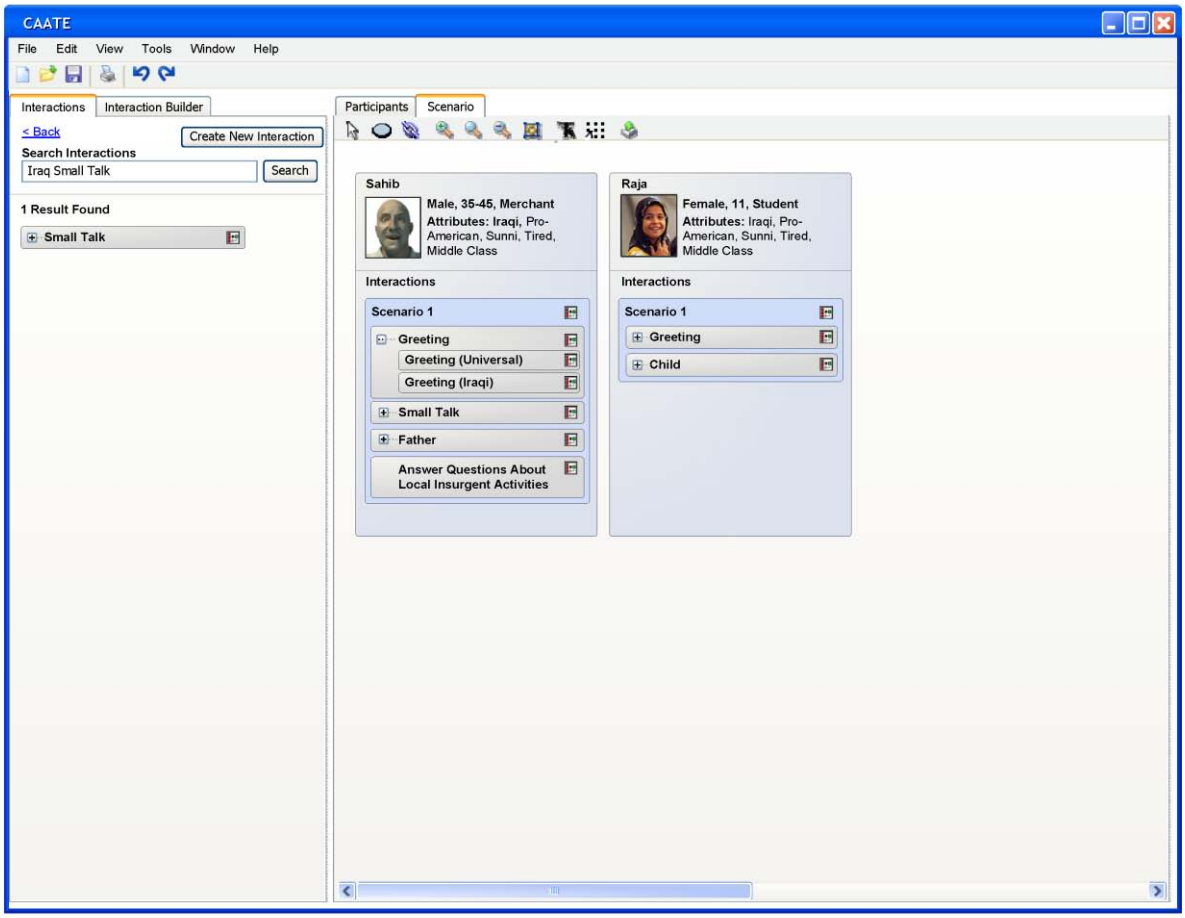


Figure 7. Behavior manager.

Defining what these behaviors are, however, does not tell the system how they are to be carried out or how they interact with each other. Figure 8, which is based on our in-house AgentWorks agent development tool, shows how these behaviors are defined and modified. In this example, the developer is editing the Iraqi Greeting behavior where the agent listens to the Soldier's greeting, responds appropriately, and then waits for the Soldier to continue the conversation. If, however, the Soldier does not initiate the greeting, the agent will. If the developer were to edit Sahib's "Greeting" behavior, the two sub-behaviors would be the boxes on the workspace for editing. In this case, both sub-behaviors would be active, though the developer would specify that when the sub-behaviors produced conflicting behaviors (if, for instance, the Iraqi Greeting included Iraqi-specific proxemics behaviors), the more specific Iraqi Greeting behavior should take precedence.

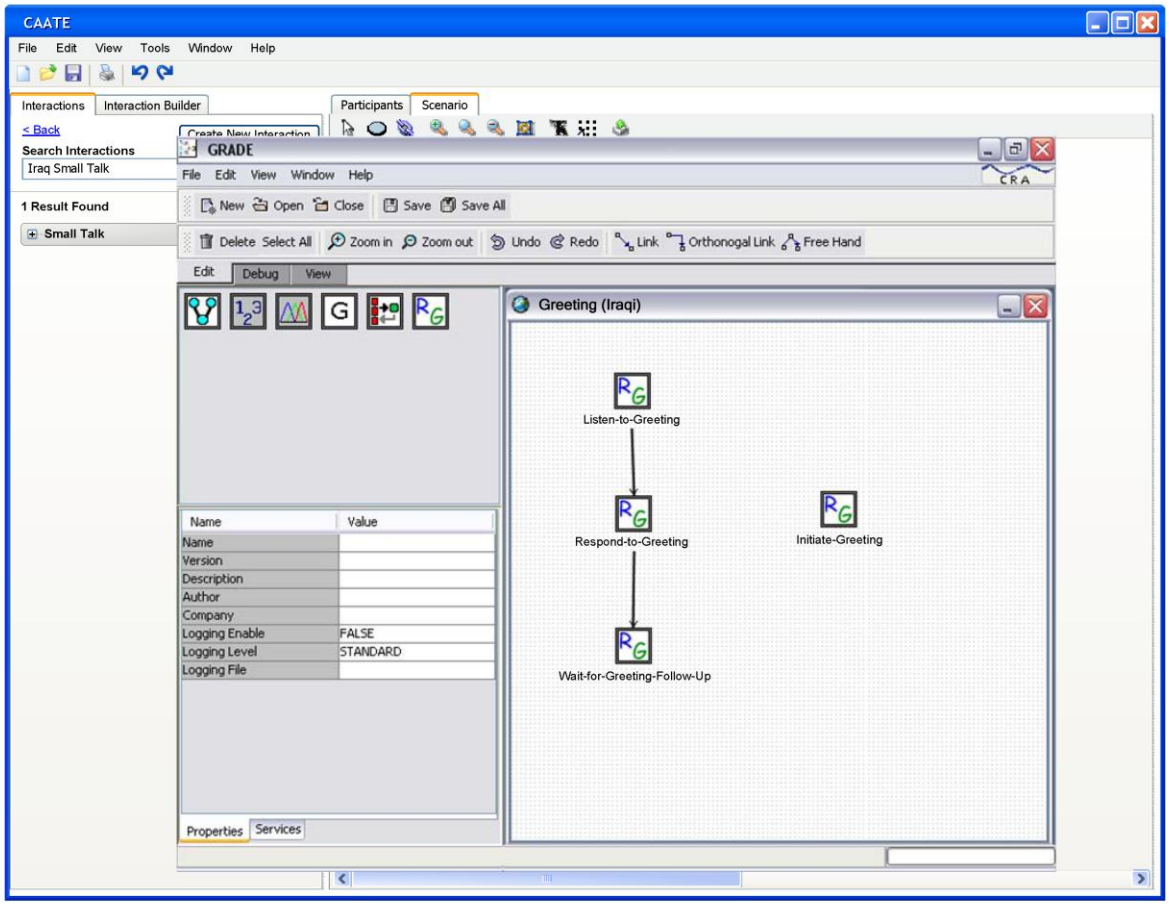


Figure 8. Creating/editing a behavior.

In Figure 9, we see the developer drilling down one step further and editing one of the component behaviors. In this case, the developer is assigning appropriate greetings based on any relevant social or cultural cues. In this example, another component has computed a "Willingness to Interact" variable based on features such as the relationship with the person being greeted and the agent's personality and emotional state. If the agent is positively inclined towards interaction, it will say "Greetings." Otherwise, it will take a less communicative option.

Design and Demonstration of Modular CAATE Runtime Architecture

During Phase I, we designed and prototyped a simple technical demonstration of a three-agent scenario that uses all of the aspects of a CAATE VE. We have used the Half-Life 2 game platform to model the physical world and agents. Figure 10 shows the runtime architecture that we have designed and prototyped for the CAATE system.

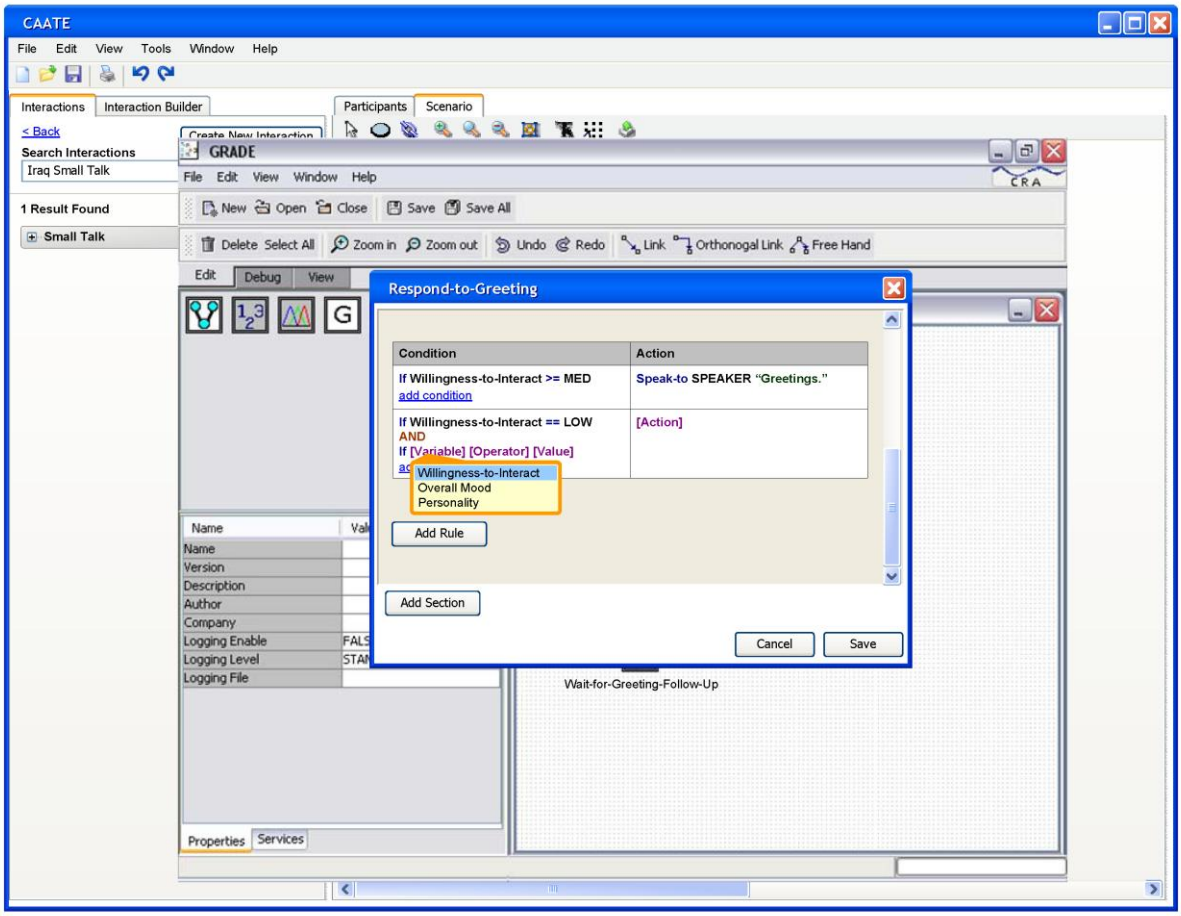


Figure 9. Editing behaviors with cultural and social moderators.

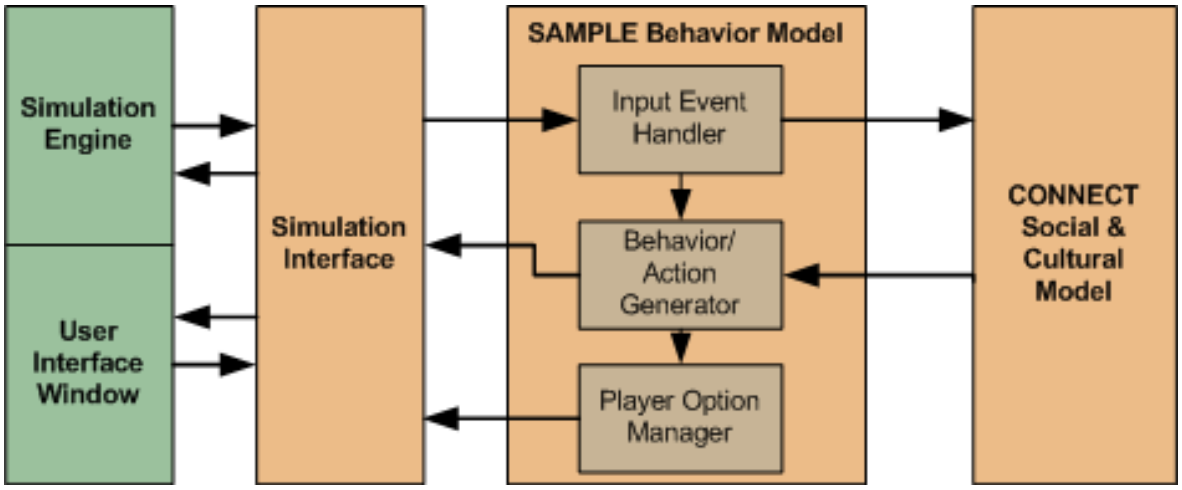


Figure 10. CAATE runtime architecture.

Half-Life 2 was our Phase I *Simulation Engine*, though we have designed the system to be independent of any particular simulation engine. The *Simulation Interface* provides a modular interface between the CAATE agents and the simulation engine, making it possible to port from one engine to another without having to modify the agents.

CAATE also provides a *User Interface Window* (see the right-hand window in Figure 11) that handles any inputs or outputs that are not directly handled by the simulation engine. For instance, if the simulation engine does not provide a means (or provides an expensive means) for displaying facial expressions, then facial expressions can be described by the user interface window. We discuss this feature in more detail below.

The CAATE agents consist of two main parts, the *SAMPLE Behavior Model* and the *CONNECT Social and Cultural Model*. The behavior model is the main interface to the simulated environment. Sense data come into an *Input Event Handler* which passes them on to CONNECT, which updates the current state of the cultural moderators (e.g., relationships, emotions) based on the sensed events and actions. The *Behavior/Action*

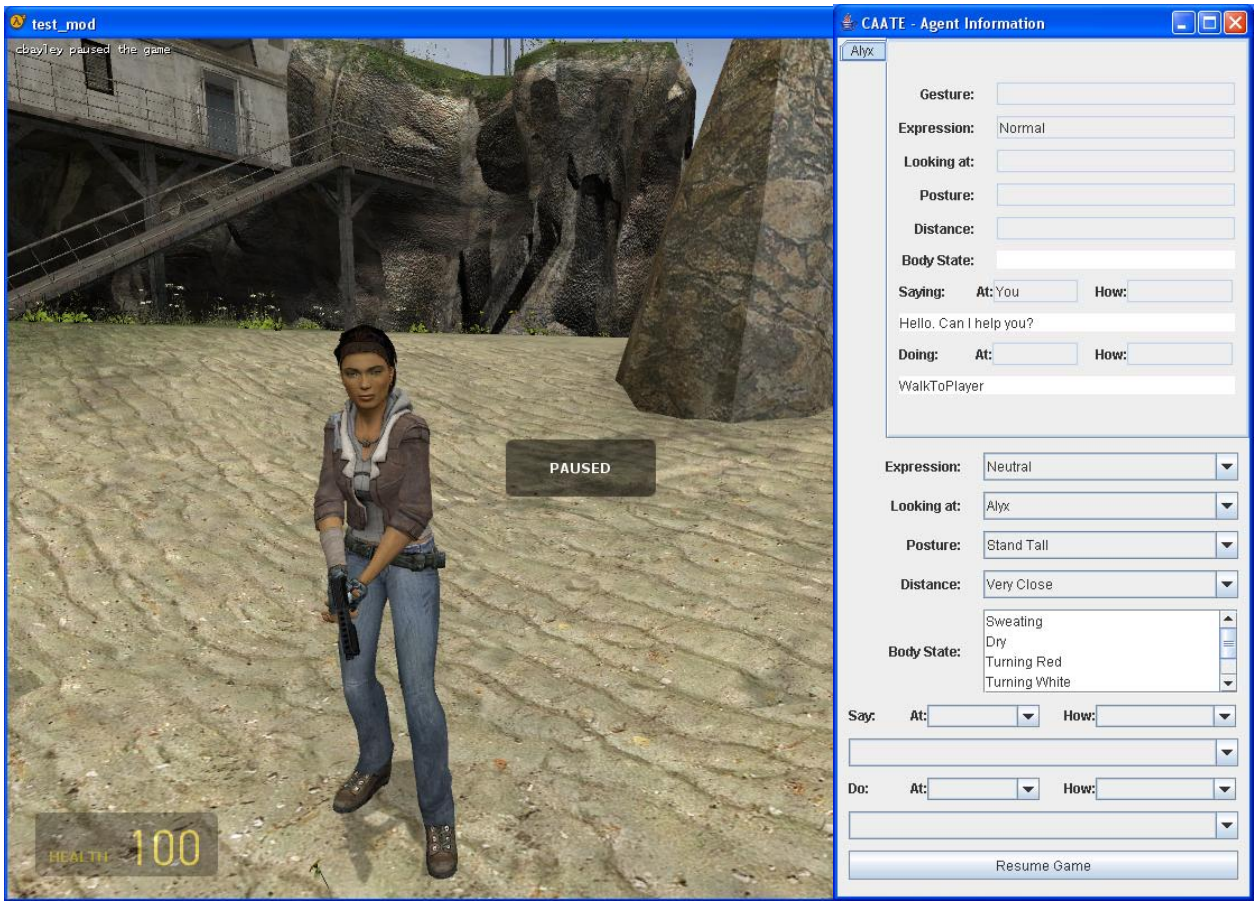


Figure 11. CAATE demonstration prototype: simulation and user interface window.

Generator chooses how to respond, based on the sense data and on the updated social/cultural state. The agent's actions are passed to the simulation interface, which dispatches them to the simulation or the user interface window as appropriate. The agent's actions also open up a number of plausible trainee responses. The *Player Option Manager* is informed of the chosen action and sends a set of plausible responses to the simulation interface, which updates the user interface window as needed. For instance, if the agent asks a question, the trainee may chose to say "yes," "no," or any of a number of other responses. The interface will add these choices to the list of items on the trainee's drop-down menu of possible speech acts to take.

One of the objectives of the CAATE effort is to ensure that the resulting system is portable across simulation platforms. We have designed the interface between the CAATE agents and the Half-Life 2 simulation platform with portability in mind. One key aspect of this portability relates to the basic input and output capabilities of different simulation engines. Half-Life 2, for instance, provides support for facial animations. Many simulation platforms do not provide such support (or do so in a way that makes it infeasible to use for many developers). In these systems, we still need the agents to "display" facial expressions. To support a wide range of simulation engine capabilities, we have developed an I/O interface window that provides the trainee the means of behaving in a range of ways that the simulated agents can react to and the means of seeing what the simulated agents are doing.

Figure 11 shows this window. The agent in the simulation ("Alyx") can express her gesture, facial expression, who she is looking at, posture, distance, body state, and verbal and non-verbal actions either through the simulated environment or through this window. As most simulation platforms do not provide the expressiveness that is necessary for many training applications, this window provides the extra expressiveness. Other agents would appear as other tabs in the upper half of the window. The bottom half of the window is used by the trainee to specify his or her actions in the environment since, again, most simulation platforms do not support this range of relevant inputs. This interface provides both expressive power and portability as CAATE agents do not have to be tailored to the capabilities of the simulation platform.

Development of Agent-Based Virtual-Training Methodology

During Phase I, we began looking into how to most effectively use and evaluate the use of agent-based VEs for training social and cultural skills. In particular, we looked at how VEs could effectively be used as part of a comprehensive training regimen. For instance, non-virtual training might be used to train basic skills (e.g., image classification to teach emotion recognition), but transferring such a skill into a realistic setting with all the complexity that it entails is an excellent use of an agent-based VE.

We also looked at mechanisms for evaluating the skills of trainees. While some skills can be evaluated using off-line methods, such as questionnaires, these approaches are limited in their ability to evaluate trainee skills demonstrated during the simulation. In an effort to evaluate such skills, it is possible to extend the user interface window shown in Figure 11 to include trainee inputs that can be used to evaluate their current skills. For instance, by adding a *threat* slider, we can have the trainee choose what they believe is the current level of threat posed by the person

they are currently interacting with (due to, for instance, their anger or anti-American attitudes). By comparing this value to the current internal state of the agent, we are able to evaluate the trainee's skill in this recognition task.

In addition to these low-level evaluation mechanisms, we also identified suitable training evaluation and Verification and Validation (V&V) approaches to evaluate the degree to which CAATE-based training applications contribute to Army goals. The Army considers V&V (Headquarters, Department of the Army, 1999) to be integral to acquisition strategy, development, and life-cycle management of simulations developed after June 1992 (Headquarters, Department of the Army, 2005). We based our evaluation plan upon three philosophical approaches that have developed separately but in many ways offer parallel and complimentary suggestions for evaluating the value the Army has received from training investments.

- Computer science developed *verification, validation, and accreditation* (VV&A) processes (Headquarters, Department of the Army, 2005; Defense Modeling and Simulation Office, 2006), to assess the value of modeling and simulation software. The principal objective of VV&A processes is to ensure that modeling and simulation software products meet user needs.
- *Program evaluation*, as typified by Rossi, Lipsey, and Freeman (2004), focuses on determining the value received from investments in social programs.
- *Training evaluation*, as typified by Kirkpatrick (1998) and Kraiger, Ford, and Salas (1993), considers specifically results from training.

There appears to be some difference in how computer scientists and psychologists use the term *validity*. Computer science authors tend to discuss validity as a property imparted into a software application during development. For example, the *Verification, validation, and accreditation* r*ecommended practices guide* (Defense Modeling and Simulation Office, 2006) discusses *validation* as a step in developing software. In contrast, psychologists discuss *evidence for validity* in the context of using tests and measures. In this context, validity refers to the degree to which evidence and theory support the interpretations of scores resulting from proposed uses of tests (American Educational Research Association, et al., 1999). To psychologists, validity is not a property of an instrument or intervention, but of how it is used and its results interpreted.

Overview of verification, validation, and accreditation.

The conceptual development of V&V techniques is largely the product of software designers. The basic four terms critical to V&V are worth presenting independently.

Verification. Army Pamphlet (AP) 5-11 (Headquarters, Department of the Army, 2005) and the Defense Modeling and Simulation Office (DMSO) Recommended Practices Guide (RPG) (DMSO, 2006) define *verification* the process of determining that a model or simulation implementation accurately represents the developer's conceptual description and specifications in the requirements document. Verification also evaluates the extent to which the model or simulation has been developed using sound and established software engineering techniques, and

establishes whether the modeling and simulation (M&S) logic and code correctly perform the intended functions. In short, verification addresses the question "Have we built the model right?" M&S verification includes appropriate data verification and M&S documentation, and should be performed by an agent independent from the M&S developer.

Validation. AP 5-11 (Headquarters, Department of the Army, 2005) and the DMSO RPG (DMSO, 2006) define *validation* as the process of determining the degree to which a model or simulation is an accurate representation of the real world from the perspective of the intended uses of the model or simulation. The validation process ranges from single modules to the entire system, with the ultimate purpose being to validate the entire system of M&S, including data. Validation considers the question "Have we built the right model?" Validation methods will incorporate documentation of procedures and results of all validation efforts to assist in the accreditation of M&S.

Accreditation. The creators of an M&S system can largely assess Verification and Validation, but accreditation is largely assessed from the perspectives of users. Accreditation is an official certification that a model, simulation, or federation of models and simulations and its associated data is acceptable for use for a specific purpose in accordance with DoDI 5000.61. It is based on experience and expert judgment and includes consideration of the extent to which V&V have been accomplished and factors that impact the decision for approval and use. Accreditation is a management responsibility of the application sponsor, assisted by the V&V agent. Accreditation answers the question "Should this thing be used?"

Credibility answers whether users trust the software for its intended purpose and see it as fit for the intended use. As with accreditation, credibility is assessed from the perspective of the users. The assessment is typically based on the experience and expert judgment of the evaluators, basically answering the question of whether the system should be trusted.

The goal of the V&V process is to determine the degree to which simulations are correct and valid, and provide simulation users with sufficient information to determine if the simulation can meet their needs. Capability and accuracy refer to whether the software does what is needed and has sufficient fidelity with intended use. Correctness is the degree of confidence that the simulation's data and algorithms are sound and robust and properly implemented, and that the accuracy of the simulation results will not substantially and unexpectedly deviate from the expected degree of accuracy. Usability consists of factors related to the use of the simulation, such as the training and experience of those who operate it, the quality and appropriateness of the data used in its application, and the configuration control procedures applied to it. Criticality refers to the costliness of consequences of errors, impact on national and military objectives and effectiveness. Criticality is related to whether intended uses are for training versus operational decision-making, and is also related to the level of simulation difficulty and technical uncertainty.

Representing human behavior in simulations. AP 5-11, the RPG, and Goerger, McGinnis, & Darken (2005) describe validation methodology for Human Behavior Representation (HBR) models. Human behavior consists of multiple, chaotic, randomly variable, non-linear, and interactive functions, making it significantly more complex to represent than the behavior of

physical objects. The interactive functions of human behavior would appear to include at least ten groups of moderators, many of which psychologists have observed and studied for years:

- Relatively stable variables that generally apply differently across groups of individuals. These variables would be exemplified by shared aspects of cultures.
- Relatively stable variables in unique patterns within each individual that use components that exist across groups. Examples of these patterns are personality characteristics, such as those of the five factor model of personality, neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness; (Goldberg, 1993) and strongly-help personal attitudes and beliefs (Petty, 1995). The connectionist paradigm (Bechtel & Abrahamsen, 1991; Read & Miller, 1998) offers potential to understand how humans process information.
- Relatively unstable variables of individuals and groups, as exemplified by influences of current environment and situation; current stress, workload, and fatigue; and weakly-help attitudes and beliefs. The strength of situational cues is typically under-recognized (Gilbert, 1995).
- Highly unstable cognitive processes resulting from current thought, attention, and perception. For example, in varying circumstances, the same individual can fill varying roles as spouse, leader, follower, employee, parent, or group member, which can lead to substantial differences in observable behavior across situations despite stable personality characteristics.
- Reaction and interaction with observers and other actors. Simply observing behavior can change it, as was first observed in the classic studies of production workers at the Hawthorne Electric Works (the *Hawthorne effect*) (Whitehead, 1938). In well-structured social situations, behavior can be governed by *scripts* (such as the restaurant script for ordering a meal) (Graesser, Woll, Kowalski, & Smith, 1980) that lead to consistent behavior regardless of the specific actors involved. Conversely, in less structured social situations, behavior can be different, depending on the individuals involved. Dynamical Systems Theory (DST) (Vallacher & Nowak, 1994; 1997) offers insight into the complexities inherent in interpersonal interaction. As one example, humans make attributions about whether someone else's behavior is due to situation or disposition, and adjust their perceptions and behavior accordingly (Gilbert, 1995). They also use a wide variety of techniques to lead (Bennis, 2007; Lord, 2001) and influence others (Cialdini, 1995). During conflict, opponents typically attempt to take advantage of these processes to disguise their true intentions and activities (Dewar, 1989).
- Group effects, in which outcomes differ from simple multiplicative values of (individual capacity * number of workers * time). An example of one such phenomenon is *social loafing*, first reported by Latané, Williams, and Harkins (1979), in which all members of a group do not contribute at their equivalent individual levels of effort.
- Learning, in which humans acquire expertise. Humans learn from experience and then use the learning to modify future responses to environmental stimuli (Pew & Mavor, 1997). Each individual's mental and physical abilities place limitations on his or her ability to acquire knowledge and skills, and learning is itself moderated by motivation and opportunities to learn.

- Much human behavior may not seem rational or under conscious control. It is clear that humans are not always aware of all reasons why they are doing things (Kihlstrom, Barnhardt, & Tataryn, 1992), particularly in focusing attention and awareness (Jacoby, Lindsay, & Toth, 1992) and acquiring information (Lewicki, Hill, & Czyzewska, 1992).
- Further, cultural differences and differing goals may lead to behaviors that observers would consider irrational. For example, understanding the Iraqi regime's values and priorities suggests that their responses to Operation Iraqi Freedom were not so poorly planned as Americans thought at the time (Woods, Pease, Stout, & Lacey, 2006).

Pew and Mavor (1997) note that simulations of human behavior require a number of components to accurately mimic the processes that underlie behavior. These include sensing and perception, working memory (the classic 7 + or – 2 chunks of information at one time) (Miller, 1956), long-term memory, situation assessment, decision making, task management, and motor response.

Overview of training and program evaluation.

Program evaluation, as typified by Rossi, Lipsey, and Freeman (2004), focuses on determining the value received from investments in social programs. Social programs are organized efforts to reduce or eliminate a social problem or improve social conditions. Program evaluation aims to distinguish between effective and ineffective programs and to provide guidelines to achieve the desired results when creating new programs or redesigning existing ones. Factors typically included in a program evaluation include the nature and scope of the problem; justification for dealing with it; feasibility of creating interventions and reaching appropriate target populations; efficiency and effectiveness in implementing interventions; whether the interventions are achieving stated goals; and comparison of costs and benefits.

Training evaluation, our final paradigm, focuses specifically on results obtained from training. The most frequently used training evaluation approach, first proposed over 40 years ago by Donald Kirkpatrick (1998), considers: (1) Trainee *reactions*, essentially customer feedback from training participants; (2) *Learning*, most often assessed by end of instruction tests; (3) *Behavior*, which is operationalized by on-the-job behavioral changes; and (4) *Results*, which are most often operationalized as organizational benefits such as reduced costs or increased output. Phillips (1997) added a fifth level that weighs costs and benefits to determine *return-on-investment*. Similarly, Cascio (1989, 1991) provided guidelines for determining costs and benefits of organizational programs, including training investments (see also Raju, Burke, & Normand, 1990). Despite its widespread use, critics, such as Alliger and Janak (1989), have noted shortcomings and misunderstandings in the Kirkpatrick paradigm. Kraiger, Ford, and Salas (1993) proposed a training evaluation approach that includes three categories of outcomes: cognitive gains, skills and performance enhancement, and adjustments to attitudes and motivation. It should also be noted that utility and return-on-investment analysis has proven beneficial in evaluating the training activities of commercial organizations, government applications frequently provide no feasible way to establish viable financial measures of the most critical organizational outcomes. For example, it is difficult to estimate a dollar value of preventing a terrorist attack. We propose to blend the approaches suggested by Kirkpatrick (1998) and Kraiger et al. (1993), including relevant features from each.

Proposed VV&A and Evaluation Steps.

Step 1: Define the problem and establish objectives. The first step of our evaluation will define the problem, identify what features and characteristics an acceptable solution must deliver, and establish training objectives. Problem definition is critical to a successful solution. For complex problems, a formal problem analysis can provide the guidance needed to select appropriate methods and establishes a firm foundation upon which the rest of the overall process can build.

The problem statement will identify key issues to be resolved, objectives to be met, and establish the scope, conditions under which the problem should be addressed, and what characteristics of the user domain need to be considered. The problem and objectives will be described clearly enough that decisions can be made about how to solve the problem and set requirements for aspects, features, conditions or characteristics that need to be addressed in the solution. The problem and acceptable solution statements will answer questions concerning capabilities the model or simulation will require, decisions that will be made on the basis of the simulation, ramifications and consequences of errors, and risks if erroneous results are accepted. It will describe acceptability criteria to determine when success has been achieved.

At the same time, we will review progress to verify completion of relevant steps in the instructional system design (ISD) process (Tennyson, 2000; Tennyson, Schott, Seel, & Dijkstra, 1997). The ISD process calls for similar activities to specify the problem and determine the range of acceptable solutions. One specific requirement will be to develop training objectives, which are needed to guide assessment of student learning in later steps. Optimally, behavioral learning objectives include three pieces of information: The desired behavior, standards for performance to be considered as successful, and a brief description of relevant factors in the situation in which the behavior must be produced. For example, a behavioral learning objective for changing a flat tire could read: "Given a passenger car with a flat tire, replace the flat tire with the spare within 30 minutes while following all steps and precautions listed in the owner's manual."

The first step will also develop training objectives. Kraiger et al. (1993) provide three high-level outcomes for training. These are (a) to provide gains in cognitive outcomes (verbal or declarative knowledge), (b) skills, and (c) attitudes and motivations. These outcomes can be measured at individual or higher levels.

Cognitive outcomes, which are part of ability, consist of gains in verbal (or declarative) knowledge, in how knowledge is organized, and in acquisition of cognitive strategies to access, apply, and exploit the gains in knowledge. Cognitive gains include the acquisition of improved mental models, self-insight, and metacognitive skills, and can be measured as increases in amount of knowledge, accuracy and speed of recall, similarities to an ideal understanding of a subject or topic, comprehension of interrelationships and hierarchical ordering among knowledge elements, and enhanced self-awareness of knowledge gains, capabilities, and limitations.

Skill-based outcomes, which are a second part of ability, include skill compilation and automaticity (Logan, 2005; Moors & De Houwer, 2006). When skills are characterized by high

automaticity, there is fast reaction time in producing behavior that is reliable and repeatable without conscious attention or allocation of mental resources. Because non-automatized processes require allocation of mental capacity, processes that do not require capacity cannot be controlled or limited by the allocation of capacity. Thus, automatized processes are not subject to interference from concurrent tasks because task interference occurs only between processes that compete for mental capacity. Training that produces automaticity provides practice and application to produce skill-based outcomes such as improved speed and fluidity of performance, lower error rates, chunking and compiling of procedural steps, generalization of skills from one application to related applications, and reduction in requirements for conscious attention to individual procedural steps.

Attitudinal and motivational outcomes include manipulations of direction and strength of outlooks, positions, or points of view toward objects and willingness to perform tasks. While cognitive and skill-based training outcomes provide ability to perform tasks, training that targets attitudinal and motivational outcomes aims to enhance willingness to perform them.

It could also be useful to consider the various levels within the Army and between the Army and outside individuals and organizations. Examples of the various levels would include individual-level outcomes, such as test scores of individual Soldiers after they complete training. It would also be possible to evaluate unit-level outcomes, such as by comparing attitudes and confidence of a specific infantry unit that has completed the training to one that has not. It would also be possible to evaluate higher-than-unit-level outcomes, such as attitudes of Iraqi civilians toward encounters with units that had completed training compared to those who did not, or overall attitudes of Soldiers about their interactions with Iraqis over time.

Step 2: Describe the range of what constitutes an acceptable solution. The second step of the evaluation process will be to narrow the field of potential problem-solving approaches. While the nature of this contract has determined that modeling and simulation is appropriate, we may need additional information about factors that would refine the desirable characteristics of the final product. These factors include the degree to which important aspects of the real world must be recreated as if the actual event or operation were taking place; whether an event or operation must be replicable under controlled conditions; whether the software should be capable of compressing time for the important aspects of events or operations. This step will verify that requirements for the simulation match those needed for the current problem and are correct, consistent, clear, and complete.

Step 3: Define Roles and Responsibilities. The RPG uses the term *user* to represent the organization, group, or person responsible for the overall application. The simulation is created to meet the needs of its users to solve a problem or make a decision. Therefore, users should be heavily involved in defining the requirements, establishing the criteria by which acceptability will be assessed, determine what method or methods to use, making the accreditation decision, and ultimately accepting the results. Users will determine level of accreditation: Full, Limited or conditional, Modification of the simulation is needed, Additional information is needed, or No accreditation. In defining the problem, users should first identify the issues involved and establish the objectives that have to be met to solve the problem. This can be done by addressing some basic issues such as; the basic problem to be solved, particular aspects of the problem that

the simulation will help solve, defining the scope of the to-be-solved problem and how the boundaries or mission space apply, determining what decisions will be made based on the simulation results, and assessing risks that might result from acceptance of erroneous simulation outputs. In all of these steps, users must function as the needed *SME*.

The M&S Program Manager (PM) plans and manages resources for simulation development, directs the overall simulation effort, and oversees configuration management and maintenance of the simulation. The PM identifies the sources of greatest risk to the development effort and should work to control them as much as possible. Specific responsibilities include identifying the development paradigm, in coordination with the developer and directing all aspects of the development, schedule, budget, contracting, and risk management.

The *developer* actually constructs the simulation, prepares data for use in the simulation, and provides technical expertise regarding simulation capabilities as needed by the other roles. The developer creates and executes the development plan in coordination with the PM, including identifying the development paradigm, allocating resources, and establishing the schedule. He or she defines the simulation domain requirements in coordination with the user; identifies and prepares data needed to develop and execute the simulation; develops the conceptual model based on the requirements of the application; designs, implements, tests, and integrates the software, and is responsible for ensuring the simulation is built to meet user objectives and requirements. The evaluation process supports the developer by helping to reduce development risk and increase credibility. Developers can assist in some evaluation developer with timely information regarding issues and problems, allowing them to be resolved before they have a major impact on the development process.

Typical developer responsibilities associated with specific evaluation tasks include; participating in the development and execution of the evaluation plan, supporting event coordination and communication (e.g., notifying the evaluation Agent of development reviews and the availability of products required for review), collecting and storing information, coordinating development and evaluation activities, ensuring planning and scheduling are coordinated so the evaluation effort can run concurrently, and participating in the requirements definition to ensure they evolve as needed for the application. In direct support of V&V, developers:

- Provide access to data used in the development, testing, and execution of the simulation.
- Ensure development products provide the proper artifacts for the V&V activities.
- Conduct developmental tests and collect verification data.
- Integrate developmental and operational testing with the V&V effort to optimize resources.
- Work with the PM and V&V Agent to make tradeoffs between development resources and V&V resources.
- Ensure V&V reports are reviewed by participants in a timely manner.
- Establish procedures for correcting problems identified by the V&V process in a timely manner.

- Provide support as required for the accreditation assessment.

The V&V Agent plans and performs V&V activities to providing evidence of the simulation's fitness for the intended use. As discussed above, the major objectives of the V&V effort are to ensure that the simulation being developed meets the needs of the intended use, reduce development and operational risk of the simulation, enhance the simulation's credibility, and support the simulation's accreditation for the intended use.

The Accreditation Agent is the term used to define the role responsible for conducting the accreditation assessment. The Accreditation Agent provides guidance to the V&V Agent to ensure that all the necessary evidence regarding simulation fitness for use is obtained; collects and assesses the evidence; and, provides the results to the User, the role with the responsibility of making the accreditation decision (i.e., accreditation authority).

The primary objective of the Accreditation Agent is to prepare for and conduct a cost-effective accreditation assessment that results in a logical, sufficient, and fully justified accreditation recommendation to the user. Accreditation is a judgment that a simulation is fit for a specific purpose. The Accreditation Agent plans and performs the accreditation assessment and assists the user with activities that help establish the scope of the problem to be addressed. The Accreditation Agent serves as the user's advocate throughout the development process to ensure that the simulation will meet the user's requirements and that sufficient evidence is available to justify an accreditation decision.

Step 4: Develop V&V Plan. Development of the V&V plan will include identifying objectives, priorities, tasks, and products of the V&V effort; establishing schedules; allocating resources; etc. in coordination with simulation development and accreditation plans. Data V&V activities should complement the different development phases and V&V activities. The V&V Agent should work closely with the Accreditation Agent to identify data-related assessment priorities and appropriate acceptability criteria. This work facilitates the selection of data V&V activities most suited for providing evidence to support the accreditation decision.

The affinity between model algorithms and their associated data will be of primary concern because of the direct impact such affinity has on simulation credibility. Appropriateness and sufficiency of all data associated with the simulation (reference, hard-wired, and instance) will be considered in verification planning. Planning the data validation effort will also include identifying appropriate validation activities and expected outcomes as well as identifying and evaluating appropriate validation to be used in the results validation. Validation planning will also address the impact of obtaining and evaluating validation data on development program timelines and resources.

Step 5: Validate Conceptual Model. Validation of the conceptual model will develop evidence to demonstrate that the capabilities indicated in the conceptual model embody all the capabilities necessary to meet the requirements. The conceptual model should adequately specify both physical and behavioral aspects of the problem domain and appropriately traces operational requirements in the emerging design. Data availability and data appropriateness are key considerations during this phase because of their impact on model design. Several data-related

tasks that can be done during this phase include verifying data sources and availability, adequacy of metadata, input databases, output data, and developing validation data. In simulation, it is virtually impossible to separately evaluate a model and the data it uses. This is because it is the interaction of data and code that produces simulation results make both responsible for simulation credibility. This mutual dependency suggests that data V&V activities should be considered part of the overall V&V process. Indeed, data V&V activities are discussed as part of the V&V process throughout the RPG. However, because of the large number of data categories used in a simulation and the amount of time needed to locate and acquire individual data sets, data V&V has a very unique nature.

The impact of the input data on the performance of individual components and on the integrated simulation will be assessed. For each model validation test, key data elements should be tracked to ensure appropriate output. Sensitivity excursions can be run to test boundary conditions on key data elements to assess the impact of data ranges on model output. Data validation can also be conducted incrementally. For example, the terrain database for a battle simulation can be validated before battle entities and objects are added.

Data validation is performed to ensure that input data are appropriate for use in a particular simulation for a specific application. All data used to drive a model are subject to validation; but because the quantity of data may make this impractical, there may be a need to identify and prioritize key data components that most directly impact the performance of the model for the application.

Discrepancies between simulation outputs and the validation data will be examined to determine probable cause (code, input data, output data, validation data, or a combination). The divergent output should be retraced through the code, key algorithms, and input data. When the culprit has been identified, the information is recorded and recommendations made to eliminate the problem.

Step 6: Verify Design. Design verification will develop evidence to demonstrate that the design is faithful to the conceptual model, and contains all the elements necessary to provide all needed capabilities without adding unneeded capabilities. The focus of design verification is to ensure that all features, functions, behaviors, and interactions defined in the design can be traced back to the requirements expressed in the conceptual model and that all requirements expressed in the conceptual model are articulated in the design. The primary data-related V&V activities associated with design verification are described in the paragraphs below.

Step 7: Verify Implementation. The step to verify implementation will develop evidence to demonstrate that the code is correct and is implemented correctly on the hardware. Requirements will be traced to the implemented software components, individual algorithms and components are tested to ensure that they perform as designed, and data/code relationships are reviewed for appropriate operation.

The initialized data sets (i.e., the aggregated sets of transformed input instance data, in their initialized or start-up state) are checked to ensure that they continue to correspond to the

original data, have been transformed as intended, and have maintained the accuracy, fidelity, and integrity required for the intended use.

Hard-wired data will be evaluated separately because they typically consist of individual fixed constants used in specific algorithms or formulas. They will be validated along with the algorithms into which they are placed and checked by executing the associated individual algorithms. Deviations will be assessed to determine the cause (i.e., statement or execution of the algorithm, hard-wired data, or validation data) and recommendations made to resolve the problem.

The code implementing individual algorithms and models will be examined to ensure that these algorithms and models provide output data to support the needs of the application. This review should include data characterization (e.g., fidelity, format, completeness) as well as methods of collection and preparation.

Results validation determines the extent to which the simulation addresses the requirements of the application. Because the data and the simulation are inextricably intertwined (i.e., if one is not valid, then the validity of the other cannot be demonstrated), their validations are usually conducted in concert. This activity examines the extent to which the simulation, driven by valid input instance data, provides appropriate responses when exercised in the context of the application. Beginning in the implementation phase, individual components or modules are executed to ensure appropriate performance and output. Additional testing is done as the components are integrated. Ultimately, the complete simulation is executed and the resulting data are compared to the validation data to determine if the simulation is producing credible, appropriate answers.

Step 8: Evaluate Results. At this point, we will develop evidence to demonstrate the degree to which the simulation addresses the requirements of the intended use in altering behavior of Soldiers interacting with Iraqi civilians. Behavior results from an interaction between ability, motivation, and opportunity. All three circumstances are necessary and sufficient to produce behavior, that is, for behavior to occur, an individual must be able and willing to produce the behavior, plus have an opportunity to perform it.

Cognitive outcomes can readily be measured through the use of pencil-and-paper or computer-based tests that measure recognition and recall, breadth of knowledge, and speed in applying knowledge. It may be feasible to build progress checks into the simulation software. Multiple-choice format measures should use three alternatives (Sidick, Barrett, & Doverspike, 1994).

For CAATE, skill-based outcomes could be assessed through behavioral observation, hands-on testing, structured interviews, and embedded measurement. Measures of skill changes should be built into the software so as to be transparent to users.

For CAATE, attitudinal and motivational outcomes could best be assessed by using self-report measures of attitude and motivational strength, self-efficacy (confidence in ability to perform the behaviors), levels of goals set for individual performance.

The training should also include a measure of the Kirkpatrick (1998) level of trainee reactions (which are essentially customer feedback from training participants) at least at the end of the training. Including multiple levels in evaluating training effectiveness can help improve transfer of training to on-the-job activities and help to understand differing needs at varying levels (Kozlowski, Brown, Weissbein, Cannon-Bowers, & Salas, 2000). The Kirkpatrick (1998) level of learning would be included within the Kraiger et al. (1993) outcome measures, and therefore would not require a separate measure.

Assessment of the Kirkpatrick (1998) level of behavior, which is typically conceptualized as on-the-job behavioral changes, would optimally require measurement at some follow-on period once students have returned to their duty stations or been deployed. Ford, Smith, Sego, Quiñones (1993) provided a useful paradigm for this type of investigation, in which graduates are surveyed to ask about the degree to which they have performed on-the-job tasks that are analogous to the learning objectives of the training. For example, if a learning objective was set for application of information about interaction with females, the survey would ask how often the graduate had needed to apply this information and how useful the training was in dealing with situations of this type.

Assessment of the Kirkpatrick (1998) level of results is typically operationalized as organizational benefits such as reduced costs or increased output. These measures may be available from archival sources, or can be collected using survey methodology from commanders, managers, or administrators.

Step 9: Accredit the Simulation. Accreditation is the official certification that a simulation and its associated data are fit for use in the specified application. The first step will be to develop the accreditation plan. The accreditation plan should identify all the information needed to perform the accreditation assessment and their priorities, tasks, schedules, participants, etc., in coordination with simulation development and V&V plans. The information needed for the assessment is collected from the V&V effort and other sources and evaluated to determine its completeness. The fitness of the simulation is assessed using all the evidence collected from the V&V effort and other sources, and an accreditation report and recommendations are prepared for the user.

Although accreditation is often perceived as occurring at the end of a development process, the actual assessment process should begin as early as possible so V&V activities and testing activities can be sure of providing appropriate and sufficient information to support the accreditation decision. The accreditation decision is essentially the user's belief in the credibility of the simulation. The V&V effort and the accreditation assessment amass evidence to show what risks are associated with using the simulation and how likely or unlikely they are to occur. The user weighs the risks against the evidence of the simulation's capabilities. Accreditation can produce five results:

- *Full accreditation*. The simulation produces results that are sufficiently credible to support full application.
- *Limited or conditional accreditation*. Constraints limit how the simulation can be used.

- *Modification of the simulation is needed*: Modifications and subsequent V&V are needed to correct deficiencies.
- *Additional information is needed*: Not enough information for either full or conditional accreditation.
- *No accreditation*: The simulation does not adequately support the application.

Design of Evaluation Plan

During Phase I, we investigated and developed methodologies and plans for effectively evaluating the success of CAATE and the training applications built with CAATE. Based on this analysis, we recommend a two-part evaluation methodology. First, it is useful to evaluate the tools themselves in terms of whether they provide the key features described above (i.e., modularity, flexibility, reuse of key content, key elements of cultural expressiveness) and whether they are efficient enough to be used on modern desktop development machines.

Second, it is useful to evaluate the effectiveness of the training applications built with the CAATE tools. During Phase I, we developed an initial methodology for evaluating the effectiveness of simulation-based cultural training applications based on existing training evaluation methodologies. This methodology is described in the Development of Agent-Based Virtual-Training Methodology Section. For follow-on work focused on evaluating CAATE-based training applications, we recommend using a variant of this methodology that uses a combination of mechanisms to effectively evaluate CAATE training applications within the budget and schedule constraints of a Phase II SBIR effort. In particular, we recommend using a combination of subjective validation with training experts, evaluations with cultural experts, and small-scale evaluations with trainee subjects. Training experts will be able to provide feedback on our training methodology and implementation. Cultural experts can judge whether the agents act appropriately and whether the scenario is believable. In particular, if cultural experts are not able to successfully navigate the cultural encounter, that will be significant evidence that our agents and scenario are not culturally plausible. Evaluations with trainee subjects will be arranged in conjunction with the sponsor, possibly using students from a local Boston/Cambridge university, students from the Univ. of Central Florida, or students from West Point.

Conclusion

Under the Phase I effort, we designed and demonstrated the feasibility of CAATE, a tool for the affordable development and deployment of culturally believable agents for simulation-based training applications. After evaluating the functional requirements for simulation-based cultural training system for the U.S. Army and the tools to support them, we designed a simple training scenario that could be used to demonstrate and evaluate our approach during Phase I. Then we developed a set of cultural dimensions and behaviors that Soldiers need to be able to understand in order to be effective in various culturally situated operations. We designed a set of tools for building software agents that demonstrate rich, believable cultural dimensions and corresponding behaviors. We also performed an initial investigation into potential training methodologies that can be effectively employed when using virtual, simulation-based systems

for training cultural skills. We also investigated means for evaluating the CAATE methodology and tools during Phase II. Finally, we developed a simple runtime prototype that demonstrated the technical feasibility of integrating CAATE agents into simulated environments in a modular, portable manner.

Based on our Phase I results and the solicitation objectives, we recommend a Phase II effort that focuses on the development and evaluation of a full-scope CAATE prototype, a tool for the affordable development and deployment of culturally believable agents for simulation-based training applications. We recommend the following specific objectives for such an effort:

- Develop a deeper understanding of the critical cultural dimensions and behaviors that Soldiers need to be able to recognize and react to for operational success.
- Develop a simulation-based training methodology that guides the training and evaluation of social and cultural skills in virtual environments that incorporate computer-controlled social agents.
- Develop and demonstrate the effectiveness of the CAATE methodology and development tools for creating culturally believable agents that can be part of a training regimen that improves critical social skills for Soldiers in operational settings.
- Demonstrate that the CAATE methodology and tools apply to a variety of culturally situated, non-kinetic operational training objectives.

The successful completion of these objectives during a phase II effort should result in a full-scope prototype of a cultural-agent development tool that has undergone initial user testing. This prototype can then be used as a basis for advanced development and a more thorough evaluation of a deployable system in follow-on work.

Reference List

Al-Shawi, I. M. (2006). *A glimpse of Iraq: The country, the people, and the occupation*. Self-published by author; Book and content available at http://glimpseofiraq.blogspot.com and http://www.amazon.com/Glimpse-Iraq-Ibrahim-Al-Shawi/dp/1411695186/ref=sr_1_1/002-8670301-8048810?ie=UTF8&s=books&qid=1190209740&sr=8-1

Alliger, G. M. & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. *Personnel Psychology, 42*, 332-342.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington DC: AERA.

Barsch, K., Bennett, W., Hailes, T., & Prichard, J. System Level Experiment-1. Presentation to the Air Force Scientific Advisory Board. 10-18-2007.

Bechtel, W. & Abrahamsen, A. (1991). *Connectionism and the mind: An introduction to parallel processing in networks.* Cambridge, MA: Blackwell.

Bennis, W. (2007). The Challenges of Leadership in the Modern World: Introduction to the Special Issue. *American Psychologist, 62*, 2-5.

Boot, M. (2002). *The savage wars of peace: Small wars and the rise of American power*. New York: Basic.

Boyle, W. E. (1994, Spring). Under the Black Flag: Execution and Retaliation in Mosby's Confederacy. *Military Law Review, 144*, 148-163.

Cascio, W. F. (1991). *Costing human resources: The financial impact of behavior in organizations* (3rd ed.). Boston: PWS-Kent.

Chiarelli, P. W. & Michaelson, A. G. (2005). Winning the Peace: The Requirement for Full-Spectrum Operations. *Military Review*.

Cialdini, R. B. (1995). Principles and techniques of social influence. In A. Tesser (Ed.). *Advanced social psychology* (pp. 257-282). New York: McGraw-Hill.

Clore, G. L., Schwartz, N., & Conway, M. (1994). Affective causes and consequences of social information processing. In R. S. Wyer & T. K. Srull (Eds.) *Handbook of Social Cognition* (2nd ed., Vol. 1, pp. 323-418). Hillsdale, NJ: Lawrence Erlbaum.

Costa, P. T. & McCrae, R. (1992). Four Ways Five Factors Are Basic. *Personality & Individual Differences,* 13653-665.

Defense Modeling and Simulation Office (2006). *Verification, validation, and accreditation* (VV&A) *Recommended Practices Guide* (RPG). Washington, DC: Author. Available at http://vva.dmso.mil .

Dewar, M. (1989). *The art of deception in warfare*. New York: David & Charles.

Ekman, P. (1992). An Argument for Basic Emotions. *Cognition and Emotion,* 6(3/4), 169-200.

Ekman, P. & Davidson, R. J. (1994). *The nature of emotions: Fundamental questions.* New York: Oxford University Press.

Elfenbein, H. A., & Ambady, N. (2002a). Is there an in-group advantage in emotion recognition? *Psychological Bulletin, 128,* 243-249.

Elfenbein, H. A., & Ambady, N. (2002b). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128*, 203-235.

Elfenbein, H. A., & Ambady, N. (2003a). Cultural similarity's consequences: A distance perspective on cross-cultural differences in emotion recognition. *Journal of Cross-Cultural Psychology, 34,* 92-109.

Elfenbein, H. A., & Ambady, N. (2003b). Universals and Cultural Differences in Recognizing Emotions of a different cultural group. *Current Directions in Psychological Science, 12*, 159-164.

Elfenbein, H. A., & Ambady, N. (2003c). When familiarity breeds accuracy: Cultural exposure and facial emotion recognition. *Journal of Personality and Social Psychology, 85*, 276-290.

Ford, J. K., Smith, E. M., Sego, D. J., & Quiñones, M. A. (1993). Impact of task experience and individual factors on training-emphasis ratings. Journal of Applied Psychology, 78, 583-590.

Gilbert, D. T. (1995). Attribution and Interpersonal perception. In A. Tesser (Ed.). *Advanced social psychology* (pp. 99-148). New York: McGraw-Hill.

Gilsenan, M. (2006). *Recognizing Islam: Religion and society in the modern Middle East.* New York: Tarus.

Goldberg, L. R. (1993). The Structure of Phenotypic Personality Traits. *American Psychologist,* 48(1), 26-34.

Goerger, S. R., McGinnis, M. L., & Darken, R. P. (2005, May). *A Validation Methodology for Human Behavior Representation Models*. Technical Report 1 Jan 2003-24. West Point NY: US Military Academy, Dept of System Engineering. Available at: http://stinet.dtic.mil/cgi-bin/GetTRDoc?AD=ADA433696&Location=U2&doc=GetTRDoc.pdf

Graesser, A. C., Woll, S. B., Kowalski, D. J., & Smith, D. A. (1980). Memory for typical and atypical actions in scripted activities. *Journal of Experimental Psychology: Human Learning & Memory, 6*, 503-515.

Guastello, S. J. (1995). *Chaos, catastrophe, and human affairs: Applications of nonlinear dynamics in work, organizations, and social evolution.* Mahwah, NJ: Erlbaum.

Hall, E. T. (1966). *The Hidden Dimension*. Garden City, NJ: Anchor Press/Doubleday.

Hall, E. T. (1977). *Beyond Culture*. Garden City, NJ: Anchor Press/Doubleday.

Hampden-Turner, C. & Trompenaars, F. (1997). *Riding the Waves of Culture: Understanding Diversity in Global Business*. McGraw-Hill.

Hashim, A. S. (2006). *Insurgency and counter-insurgency in Iraq*. Ithica, NY: Cornell.

Headquarters, Department of the Army (1999, 30 September). Army Pamphlet 5-11, *Verification, Validation, and Accreditation of Army Models and Simulations*. Washington, DC: Author.

Headquarters, Department of the Army (2005, 1 February). Army Regulation 5-11, *Management of Army Models and Simulations*. Washington, DC: Author.

Hofstede, G. (1980). *Culture's Consequences*. Beverly Hills, CA: Sage.

Jacoby, L. L., Lindsay, D. S., & Toth, J. P. (1992). Unconscious influences revealed: Attention, awareness, and control. *American Psychologist, 47*, 802-809.

Johnson, W. & Lester, J. (2000). Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments. *International Journal of Artificial Intelligence in Education*.

Kepel, G. (2004). *The war for Muslim monds: Islam and the West*. Cambridge, MA: Belknap.

Kihlstrom, J. F., Barnhardt, T. M., & Tataryn, D. J. (1992). The psychological unconscious: Found, lost, and regained. *American Psychologist, 47*, 788-791.

Kirkpatrick, D. L. (1998). *Evaluating training programs: The four levels* (2nd ed.). San Francisco: Berrett-Koehler.

Knapp, M. A. & Hall, J. A. (1996) *Nonverbal communication in human interaction.* Ft. Worth: Harcourt-Brace.

Kozlowski, S. W. J., Brown, K. G., Weissbein, D. A., Cannon-Bowers, J. A., & Salas, E. (2000). A multilevel approach to training effectiveness: Enhancing horizontal and vertical transfer. In K. J. Klein & S. W. J. Kozlowski (Eds.) *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions*. (pp. 157-210). San Francisco: Jossey-Bass.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology, 78,* 311-328.

Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology, 37,* 822-832.

Lewicki, P., Hill, T., & Czyzewska, M. (1992). Nonconscious acquisition of information. *American Psychologist, 47,* 796-801.

Logan, G. D. (2005). Attention, Automaticity, and Executive Control. In A. F. Healy (Ed). *Experimental cognitive psychology and its applications* (pp. 129-139). Washington, DC: American Psychological Association.

Lord, R. G. (2001). The nature of organizational leadership: Conclusions and implications. In S. J. Zaccaro & R. J. Klimoski (Eds.). *The nature of organizational leadership: Understanding the performance imperatives confronting today's leaders* (pp. 413-436). San Francisco: Jossey-Bass.

Mackey, R. R. (2004). *The Uncivil War: Irregular Warfare in the Upper South, 1861–1865*. Norman: University of Oklahoma Press.

McCrae, R. (2000). Trait Psychology and the Revival of Personality-and-Culture Studies. *American Behavioral Science,* 4410-31.

McPherson, J. M. (1988). *Battle cry of freedom: The Civil War era.* New York: Oxford.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63,* 81-97.

Moors, A. & De Houwer, J. (2006). Automaticity: A Theoretical and Conceptual Analysis. *Psychological Bulletin, 132*, 297–326.

Neal Reilly, W. S. (2006). Modeling What Happens Between Emotional Antecedents and Emotional Consequents. In R. Trappl (Ed.), *Cybernetics and Systems 2006*. Vienna, Austria.

Neal Reilly, W. S., Harper, K. A., & Marotta, S. (2007). Modeling Concurrent, Interacting Behavior Moderators for Simulation-Based Acquisition Tasks. In *Proceedings of Spring Simulation Multiconference*. Norfolk, VA.

Oatley, K. & Johnson-Laird, P. N. (1987). Towards a Cognitive Theory of Emotions. *Cognition and Emotion*, 1(1), 29-50.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions.* New York: Cambridge.

Patai, R. (2002). *The Arab mind.* Long Island City, NY: Hatherleigh.

Petty, R. E. (1995). Attitude change. In A. Tesser (Ed.). *Advanced social psychology* (pp. 195-256). New York: McGraw-Hill.

Pew, R. W. & Mavor, A. S. (Eds.) (1997). *Representing human behavior in military simulations: Interim Report*. Washington, DC: National Academy Press.

Polk, W. R. (2005). *Understanding Iraq*. New York: Harper Perennial.

Pryce-Jones, D. (2002). *The closed circle: An interpretation of the Arabs*. Chicago: Dee.

Poole, H. J. (2005). *Militant tricks: Battlefield ruses of the Islamic insurgent*. Emerald Isle, NC: Posterity.

Raju, N. S., Burke, M. J., & Normand, J. (1990). A new approach for utility analysis. *Journal of Applied Psychology, 75*, 3-12.

Read, S, J. & Miller, L. C. (Eds.) (1998). *Connectionist models of social reasoning and social behavior.* Mahwah, NJ: Earlbaum.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th ed.). Thousand Oaks, CA: Sage.

Sidick, J. T., Barrett, G. V., & Doverspike, D. (1994). Three-alternative multiple choice tests: An attractive option. *Personnel Psychology, 47*, 829-835.

Stofft, W. A. & Guertner, G. L. (1995). Ethnic Conflict: The Perils of Military Intervension. *Parameters*, 35.

Tennyson, R. D. (2000). Fourth generation instructional systems development: A problem solving approach. *Journal of Structural Learning & Intelligent Systems, 14*, 229-252.

Tennyson, R. D., Schott, F., Seel, N. M., & Dijkstra, S. (Eds.). (1997). *Instructional design: International perspectives, Vol. 1: Theory, research, and models*. Mahwah, NJ: Erlbaum.

Triandis, H. C. & Suh, E. M. (2002). Cultural Influences on Personality. *Annual Review of Psychology*, 53133-160.

Trompenaars, F. (2001). *Leaders for the 21st Century*. Capstone Publishing.

Vallacher, R. R. & Nowak, A. (1997). The emergence of dynamical social psychology. *Psychological Inquiry, 8*, 73-99.

Vallacher, R. R. & Nowak, A. (Eds.) (1994). *Dynamical systems in social psychology*. San Diego: Academic Press.

Whitehead, T. N. (1938). *The industrial worker*. (2 vols.). Oxford, UK: Harvard Univ. Press. Portions available at http://books.google.com/books?id=dXOpA0nkq8gC&pg=PA3&dq=%22the+industrial+worker%22+whitehead&sig=9H4xkW4yCvFazmeRcEKgQdkiSbU#PPA4,M1

Williams, J. E., Satterwhite, R. C., & Saiz, J. L. (1998). *The Importance of Psychological Traits: A Cross-Cultural Study*. New York: Plenum Press.

Woods, K. M., Pease, M. R., Stout, W. M., & Lacey, J. G. (2006). *A view of Operation Iraqi Freedom from Saddam's senior leadership.* Norfolk, VA: Joint Center for Operational Analysis, US Joint Forces Command. Available at http://www.jfcom.mil/newslink/storyarchive/2006/ipp.pdf

Wyer, R. S. & Srull, T. K. (Eds.) (1994). *Handbook of social cognition (2nd ed.), Vol 1*. Hillsdale NJ: Erlbaum.

Appendix A

GLOSSARY OF ABBREVIATIONS

AFRL/IF	Air Force Research Lab/Information Directorate
AOC	Air Operations Center
CONNECT	Customizable ONtology-based NEtwork Component Toolkit
ICT	Institute for Creative Technologies
CAATE	Culturally Aware Agents for Training Environments
COTS	Commercial Off-the-Shelf
GOTS	Government Off-the-Shelf
DoD	Department of Defense
SME	Subject Matter Expert
MEC	Mission Essential Competencies
VE	Virtual Environment
OOS	OneSAF Objective System
SBIR	Small Business Innovative Research
IP	Intellectual Property
KSA	Knowledge, Skills, and Abilities
NPC	Non-Player Character
I/O	Input/Output
UI	User Interface
V&V	Verification and Validation
VV&A	Verification, Validation and Accreditation
DMSO	Defense Modeling and Simulation Office
RPG	Recommended Practices Guide
M&S	Modeling and Simulation
AP	Army Pamphlet
HBR	Human Behavior Representation
DST	Dynamical Systems Theory
ISD	Instructional System Design
PM	Program Manager